STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health

Diederik Aerts
Andrei Khrennikov
Massimo Melucci
Bourama Toni *Editors*

Quantum-Like Models for Information Retrieval and Decision-Making

Springer

STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health

STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health

This interdisciplinary series highlights the wealth of recent advances in the pure and applied sciences made by researchers collaborating between fields where mathematics is a core focus. As we continue to make fundamental advances in various scientific disciplines, the most powerful applications will increasingly be revealed by an interdisciplinary approach. This series serves as a catalyst for these researchers to develop novel applications of, and approaches to, the mathematical sciences. As such, we expect this series to become a national and international reference in STEAM-H education and research.

Interdisciplinary by design, the series focuses largely on scientists and mathematicians developing novel methodologies and research techniques that have benefits beyond a single community. This approach seeks to connect researchers from across the globe, united in the common language of the mathematical sciences. Thus, volumes in this series are suitable for both students and researchers in a variety of interdisciplinary fields, such as: mathematics as it applies to engineering; physical chemistry and material sciences; environmental, health, behavioral and life sciences; nanotechnology and robotics; computational and data sciences; signal/image processing and machine learning; finance, economics, operations research, and game theory.

The series originated from the weekly yearlong STEAM-H Lecture series at Virginia State University featuring world-class experts in a dynamic forum. Contributions reflected the most recent advances in scientific knowledge and were delivered in a standardized, self-contained and pedagogically-oriented manner to a multidisciplinary audience of faculty and students with the objective of fostering student interest and participation in the STEAM-H disciplines as well as fostering interdisciplinary collaborative research. The series strongly advocates multidisciplinary collaboration with the goal to generate new interdisciplinary holistic approaches, instruments and models, including new knowledge, and to transcend scientific boundaries.

More information about this series at http://www.springer.com/series/15560

Diederik Aerts • Andrei Khrennikov
Massimo Melucci • Bourama Toni
Editors

Quantum-Like Models for Information Retrieval and Decision-Making

Springer

Editors
Diederik Aerts
Center Leo Apostel for Interdisciplinary Studies
Brussels, Belgium

Department of Mathematics
Brussels Free University
Brussels, Belgium

Andrei Khrennikov
Linnaeus University, International Center for Mathematical Modeling in Physics and Cognitive Sciences
Växjö, Sweden

Massimo Melucci
Department of Information Engineering
University of Padova
Padova, Italy

Bourama Toni
Department of Mathematics
Howard University
Washington, DC, USA

ISSN 2520-193X ISSN 2520-1948 (electronic)
STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health
ISBN 978-3-030-25912-9 ISBN 978-3-030-25913-6 (eBook)
https://doi.org/10.1007/978-3-030-25913-6

Mathematics Subject Classification: 37-04, 46C99, 60B99, 37N, 81P45, 94A05, 94A17, 94A40, 81Q93, 81Q20, 81Q65, 81S05, 81S10, 81S22

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

All of us are users of information access and retrieval (IAR) systems. Search engines for the World Wide Web (WWW), local or personal media repositories and mailbox search functions are some of the many examples that are becoming a preferred way of acquiring, aggregating, filtering and interacting with multimodal information as well as of achieving certain information goals especially when interacting with people and companies.

The peculiar difficulty of IAR is the fact that relevance cannot be precisely and exhaustively described by the data itself; for example, a relevant document may not be precisely and exhaustively described using text even if this text includes or considers all elements or aspects of a topic; a user's information need may not be precisely and exhaustively described using a query even if this query is a comprehensive description of the need.

Humans are adept at making reasonably robust and quick decisions about what information is relevant to them, despite the ever-increasing complexity and volume of their surrounding information environment. The literature on document relevance has identified various dimensions of relevance (e.g. topicality, novelty, etc.) that evolve and interact with each other.

However, little is understood about how the dimensions of relevance may interact and how this interaction is contextual and uncertain in nature. The problem becomes more complex and challenging when processing and interacting with multimodal information (e.g. linking an image with a news article, identifying regions or objects of interest within images, tagging video and music clips, etc.), due to the semantic gap between low-level multimedia content features (e.g. pixels, colour histograms, texture, etc.) and high-level meanings as well as the interference on relevance judgements for a document caused by multimodal interactions. Therefore, the current state-of-the-art of IAR is insufficient to address the challenges of the dynamic, adaptive and multimodal nature of the information and user interaction context. A genuine theoretical breakthrough is on the contrary necessary.

The quantum mechanical framework may help give up the notions of unimodal features and classical ranking models disconnected from context, thus making the emergence of quantum-like modelling of IAR possible and potentially effective at the same level of efficiency of the traditional modelling. It is believed by the authors of this volume that the quantum theoretical framework can provide the breakthrough in IAR because it can integrate abstract vector spaces, probability spaces and query languages and extend and generalise the classical vector, probability and query languages utilised in IAR.

The chapters of this volume share the aim of finding novel ways to address foundational problems that a community of researchers is actively working to define, investigate and evaluate new methods for information processing inspired by quantum theory. For many years, there exists an active research area where psychology, economy, mathematics, information science, computer science and others meet to provide and share methodological and experimental results obtained by the means of quantum theory when applied to disciplines other than physics.

In this context, the European Union funded the project “Quantum Information Access Theory” (QUARTZ) which is an Innovative Training Network (ITN) within the Horizon 2020 Marie Sklodowska-Curie Action programme. QUARTZ started from the idea of transferring the scientific research results and expertise from senior researchers in the utilisation of quantum theory in IAR to the junior researchers, thus stimulating the birth and growth of a networked European community of scholars with a larger, stronger and deeper expertise in IAR. We believe that this volume is providing an updated view of the current research in quantum-like modelling of IAR and in particular describes some of the research issues and the solutions thereof investigated within QUARTZ.

Brussels, Belgium — Diederik Aerts
Växjö, Sweden — Andrei Khrennikov
Padua, Italy — Massimo Melucci
Washington, DC, USA — Bourama Toni
April 2019

Contents

Contributors

Diederik Aerts Center Leo Apostel for Interdisciplinary Studies, Brussels, Belgium

Department of Mathematics, Brussels Free University, Brussels, Belgium

Alexander P. Alodjants ITMO University, Saint Petersburg, Russia

Igor A. Bessmertny ITMO University, Saint Petersburg, Russia

Jerome Busemeyer Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

Emanuele Di Buccio Department of Information Engineering, University of Padova, Padova, Italy

Andrei Khrennikov Linnaeus University, International Center for Mathematical Modeling in Physics and Cognitive Sciences, Växjö, Sweden

Massimo Melucci Department of Information Engineering, University of Padova, Padova, Italy

Aleksey V. Platonov ITMO University, Saint Petersburg, Russia

Massimiliano Sassoli de Bianchi Center Leo Apostel for Interdisciplinary Studies, Brussels, Belgium

Laboratorio di Autoricerca di Base, Barbengo, Switzerland

Ingo Schmitt Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany

Evgeny K. Semenenko ITMO University, Saint Petersburg, Russia

Sandro Sozzo School of Business and Centre IQSCS, University of Leicester, Leicester, UK

Tomas Veloz Center Leo Apostel for Interdisciplinary Studies, Brussels, Belgium

Universidad Andres Bello, Departamento Ciencias Biológicas, Facultad Ciencias de la Vida, Santiago, Chile

Fundación para el Desarrollo Interdisciplinario de la Ciencia la Tecnología y las Artes, Santiago, Chile

Zheng Joyce Wang School of Communication, Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH, USA

Benyou Wang Department of Information Engineering, University of Padova, Padova, Italy

Günther Wirsching Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany

Matthias Wolff Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany

Modeling Meaning Associated with Documental Entities: Introducing the Brussels Quantum Approach

Diederik Aerts, Massimiliano Sassoli de Bianchi, Sandro Sozzo, and Tomas Veloz

Abstract We show that the Brussels operational-realistic approach to quantum physics and quantum cognition offers a fundamental strategy for modeling the meaning associated with collections of documental entities. To do so, we take the World Wide Web as a paradigmatic example and emphasize the importance of distinguishing the Web, made of printed documents, from a more abstract meaning entity, which we call the Quantum Web, or QWeb, where the former is considered to be the collection of traces that can be left by the latter, in specific measurements, similarly to how a non-spatial quantum entity, like an electron, can leave localized traces of impact on a detection screen. The double-slit experiment is extensively used to illustrate the rationale of the modeling, which is guided by how physicists constructed quantum theory to describe the behavior of the microscopic entities. We also emphasize that the superposition principle and the associated interference

D. Aerts
Center Leo Apostel for Interdisciplinary Studies, Brussels, Belgium

Department of Mathematics, Brussels Free University, Brussels, Belgium
e-mail: diraerts@vub.ac.be

M. Sassoli de Bianchi
Center Leo Apostel for Interdisciplinary Studies, Brussels, Belgium

Laboratorio di Autoricerca di Base, Barbengo, Switzerland
e-mail: msassoli@vub.ac.be

S. Sozzo (✉)
School of Business and Centre IQSCS, University of Leicester, Leicester, UK
e-mail: ss831@le.ac.uk

T. Veloz
Center Leo Apostel for Interdisciplinary Studies, Brussels, Belgium

Universidad Andres Bello, Departamento Ciencias Biológicas, Facultad Ciencias de la Vida, Santiago, Chile

Fundación para el Desarrollo Interdisciplinario de la Ciencia la Tecnología y las Artes, Santiago, Chile
e-mail: tveloz@dicta.cl

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_1

effects are not sufficient to model all experimental probabilistic data, like those obtained by counting the relative number of documents containing certain words and co-occurrences of words. For this, additional effects, like context effects, must also be taken into consideration.

Keywords Quantum structures · Conceptual entities · Documental entities · Interference effects · Context effects · Information Retrieval · Word co-occurrence

1 Introduction

In his book about the geometry of Information Retrieval (IR), Rijsbergen writes in the prologue [30]:

> Well imagine the world in IR before keywords or index terms. A document, then, was not simply a set of words, it was much more: it was a set of ideas, a set of concepts, a story, etc., in other words a very abstract object. It is an accident of history that a representation of a document is so directly related to the text in it. If IR had started with documents that were images then such a dictionary kind of representation would not have arisen immediately. So let us begin by leaving the representation of a document unspecified. That does not mean that there will be none, it simply means it will not be defined in advance. [...] a document is a kind of fictive object. Strangely enough Schrödinger [...] in his conception of the state-vector for QM envisaged it in the same way. He thought of the state-vector as an object encapsulating all the possible results of potential measurements. Let me quote: ‘It (ψ-function) is now the means for predicting probability of measurement results. In it is embodied the momentarily attained sum of theoretically based future expectation, somewhat as laid down in a catalogue.’ Thus a state-vector representing a document may be viewed the same way – it is an object that encapsulates the answers to all possible queries.

In the present chapter, we adopt that part of Rijsbergen’s perspective that emphasizes the importance of distinguishing a corpus of written documents, like the pages forming the World Wide Web, made of actual (printed or printable) webpages, from the meaning (conceptual) entity associated with it, which in the case of the Web we simply call it the ‘Quantum Web’ (in short, the ‘QWeb’), because its modeling requires the use of notions derived from quantum theory, as we are going to discuss. This requirement is not at all accidental, and we are going to consider this crucial aspect too. Indeed, a strong analogy was established between the operational-realistic description of a physical entity, interacting with a measurement apparatus, and the operational-realistic description of a conceptual entity, interacting with a mind-like cognitive entity (see [13] and the references therein). In that respect, in a recent interpretation of quantum theory the non-classical behavior of quantum micro-entities, like electrons and photons, is precisely explained as being due to the fact that their fundamental nature is conceptual, instead of objectual (see [14] and the references therein). Considering the success of the quantum formalism in modeling and explaining data collected in cognitive experiments with human participants, it is then natural to assume that a similar approach can be proposed, *mutatis mutandis*, to capture the information content of large corpora of written documents, as is clear that such content is precisely what is revealed when human minds interact with said documents, in a cognitive way.

What we will describe is of course relevant for Information Retrieval (IR), i.e., [27]: "the complex of activities performed by a computer system so as to retrieve from a collection of documents all and only the documents which contain information relevant to the user's information need." Although the term "information" is customarily used in this ambit, it is clear that the retrieval is about *relevant* information, that is, *meaningful* information, so that, in the first place, IR is really about *Meaning Retrieval*. More specifically, similarly to a quantum measurement, an IR process is an interrogative context where a user enters a so-called *query* into the system. Indeed, on a pragmatic level, a query works as an interrogation, where the system is *asked* to provide documents whose meaning is strongly connected to the meaning conveyed by the query, usually consisting of a word or sequence of words. In fact, since a search engine does not provide just a single document as an outcome, but an entire collection of documents, if the numerical values that are calculated to obtain the ranking are considered to be a measure of the outcome probabilities of the different documents, the analogy consists in considering the action of a search engine to be similar to that of an experimenter performing a large number of measurements, all with the same initial condition (specified by the query), then presenting the obtained results in an ordered way, according to their relative frequencies of appearance. Of course, the analogy is not perfect, as today search engines, when they look for the similarities between the words in the query and the documents, they only use deterministic processes in their evaluations. But we can certainly think of the deterministic functioning of today search engines as a provisional stage in the development of more advanced searching strategies, which in the future will also exploit non-deterministic processes, i.e., probabilistic rankings (see [1], for an example where the introduction of some level of randomness, by means of probabilities that reflect the relative weights of the parts involved in a decision process is able to offer a more balanced way to reach a meaningful outcome; see also [28], for an explanation about how indeterminism, in measurement situations, can increase our discriminative power).

It is important to say, however, that our focus here is primarily on 'the meaning that is associated with a collection of documents' and not on the exploration of more specific properties like 'relevance' and 'information need', which are more typically considered in IR. For the time being, our task is that of trying to find a way of modeling meaning content in a consistent way, and not yet that of considering the interplay between notions like 'relevance' and 'content', or 'information need' and 'user's request' [27]. Our belief is that the adoption of a more fundamental approach, in the general modeling of meaning, will help us in the future to also address in new and more effective ways those more specific properties and their relationships.

Before entering in the description of our quantum approach, its motivations and foundations, it is useful to provide a definition of the terms "meaning" and "concept," which we use extensively. By the term "meaning," we usually refer to that content of a word, and more generally of any means of communication or expression, that can be conveyed in terms of concepts, notions, information, importance, values, etc. Meaning is also what different 'meaning entities', like concepts, can share, and when this happens they become connected, and more precisely 'connected through meaning'. By the term "concept," we usually intend a well-defined and ideally formed thought, expressible and usable at different levels, like the intuitive, logical, and practical ones. Concepts are therefore paradigmatic examples of 'meaning entities', used as inputs or obtained as outputs of cognitive activities, for instance, aimed at grasping and defining the essence of situations, decisions, reasoning, objects, physical entities, cultural artifacts, etc. Concepts are what minds (cognitive entities) are able to intend and understand, what they are sensitive to, and can respond to. They are what is created and discovered as the result of a cognitive activity, like study, meditation, observation, reasoning, etc. And more specifically, concepts are what minds use to make sense of their experiences of the world, allowing them, in particular, to classify situations, interpret them (particularly when they are new), connect them to previous or future ones, etc.

An important aspect is that concepts, like physical entities, can be in different states. For instance, the concept *Fruits*,[1] when considered in the context of itself, can be said to be in a very neutral or primitive meaning-state, which can be metaphorically referred to as its 'ground state'. But concepts can also be combined with other concepts, and when this is done their meaning changes, i.e., they enter into different contextual states. For instance, the combination *Sugary fruits* can be metaphorically interpreted as the concept *Fruits* in an 'excited state', because of the context provided by the *Sugary* concept. But of course, it can also be interpreted as an excited state of the concept *Sugary*, because of the context provided by the *Fruits* concept.

An important notion when dealing with meaning entities like human concepts is that of *abstractness*, and its complementary notion of *concreteness*. For instance, certain concepts, like *Table*, *Chair*, and *House*, are considered to be relatively concrete, whereas other concepts, like *Joy*, *Entity*, and *Justice*, are considered to be relatively abstract. We can therefore find ways to order concepts in terms of their degree of concreteness or abstractness. For example, the concept *Table* can be considered to be more concrete than the concept *Entity*, the concept *Chess table* to be more concrete than the concept *Table*, the concept *Alabaster chess table* to be more concrete than *Chess table*, and so on. Here there is the idea that concepts are associated with a set of characteristic properties, and that by making their properties more specific, we can increase their degree of concreteness, up to the point that a concept possibly enters a one-to-one correspondence with an object of our spatiotemporal theater. This is because, according to this view, concepts would typically have been created by abstracting them from objects.

[1]We will generally indicate concepts using the italic style and the capitalization of the first letter, to distinguish them from the words used to designate them. So, we will distinguish the words "juicy fruits," printed in a document, from the concept *Juicy fruits*, which such words indicate. On the other hand, words written in italic style in the article but without capitalization of the first letter of the first word are just emphasized words.

There is however another line to go from the abstract to the concrete, which can be considered to be more fundamental, and therefore also more important in view of a construction of a quantum model for the meaning content of a collection of documents. Indeed, although physical objects have played an important role in how we have formed our language, and in the distinction between abstract and concrete concepts, it is true that this line of going from the concrete to the abstract, linked to our historical need of naming the physical entities around us and define categories of objects having common features, remains a rather parochial one, in the sense that it does not take into full account how concepts behave in themselves, because of their non-objectual nature, particularly when they are combined, so giving rise to more complex entities having new emerging meanings.

When this observation is taken into account, a second line of going from the abstract to the concrete appears, related to how we have learned to produce conceptual combinations to better think and communicate (Fig. 1). The more abstract concepts are then those that can be expressed by single words, and an increase in concreteness is then the result of conceptual combinations, so that the most concrete concepts are those formed by very large aggregates of meaning-connected (entangled) single-word concepts, corresponding to what we would generically indicate as a *story*, like those written in books, articles, webpages, etc. Of course, not a story only in the reductive sense of a novel, but in the more general sense of a cluster of concepts combined so as to create a well-defined meaning. It is this line of going from the abstract to the concrete that we believe is the truly fundamental,[2] and in a sense also the universal one, which we will consider in our modeling strategy, when exploiting the analogy between a meaning retrieval situation, like when doing a Web search, and a quantum measurement in a physics' laboratory. But before doing this, in the next section we describe in some detail one of the most paradigmatic physics' experiments, which Feynman used to say that it contains the only mystery: the double-slit experiment.

In Sect. 3, we continue by providing a conceptualistic interpretation of the double-slit experiment, understanding it as an interrogative process. Then, in Sect. 4, we show how to use our analysis of the double-slit situation to provide a rationale for capturing the meaning content of a collection of documental entities. In Sect. 5, we observe that quantum interference effects are insufficient to model all data, so that additional mechanisms, like context effects, need to be also considered. In Sect. 6, we conclude our presentation by offering some final thoughts. In Appendix 1, we demonstrate that the combination of "interference plus context effects" allows in principle to model all possible data, while in Appendix 2, we introduce the notion of *meaning bond* of a concept with respect to another concept, showing its relevance to the interpretation of our quantum formalism.

---

[2]Note however that these two lines are intimately related, as is clear that one needs to use more and more concepts/words to make more and more properties describing a given situation to become more and more specific.

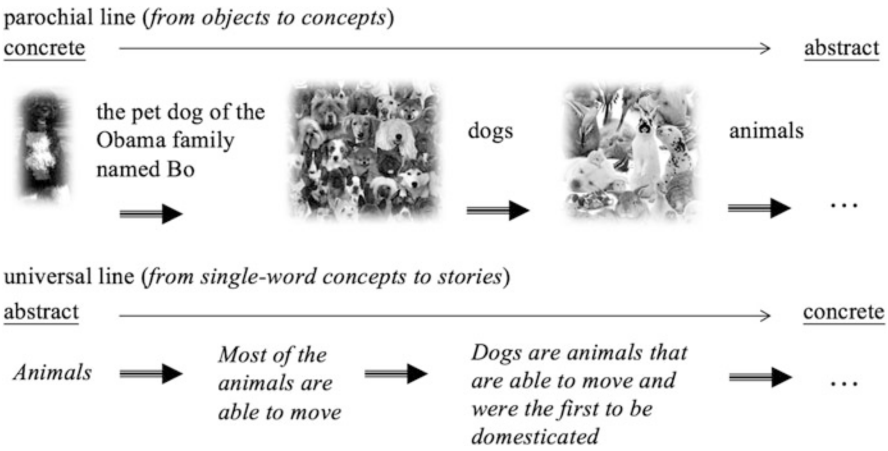


Fig. 1 Two main lines connecting abstract to concrete exist in the human culture. The first one goes from concrete objects to more abstract collections of objects having common features. The second one goes from abstract single-word concepts to stories formed by the combination of many meaning-connected concepts

2 The Double-Slit Experiment

The double-slit experiment is among the paradigmatic quantum experiments and can be used to effectively illustrate the rationale of our quantum modeling of the meaning content of corpora of written documents. One of the best descriptions of this experiment can be found in Feynman's celebrated lectures in physics [24]. We will provide three different descriptions of the experiment. The first one is just about what can be observed in the laboratory, showing that an interpretation in terms of particle or wave behaviors cannot be consistently maintained. The second (Sect. 3) one is about characterizing the experiment in a conceptualistic way, attaching to the quantum entities a conceptual-like nature, and to the measuring apparatus a cognitive-like nature. The third one is about interpreting the experiment as an IR-like process (Sect. 4).

We first consider the classical situation where the entities entering the apparatus, in its different configurations, are small bullets. Imagine a machine gun shooting a stream of these bullets over a fairly large angular spread. In front of it there is a barrier with two slits (that can be opened or closed), just about big enough to let a bullet through. Beyond the barrier, there is a screen stopping the bullets, absorbing them each time they hit it. Since when this happens a localized and visible trace of the impact is left on the screen, the latter functions as a detection instrument, measuring the position of the bullet at the moment of its absorption. Considering that the slits can be opened and closed, the experiment of shooting the bullet and observing the resulting impacts on the detection screen can be performed in four different configurations. The first one, not particularly interesting, is when both slits

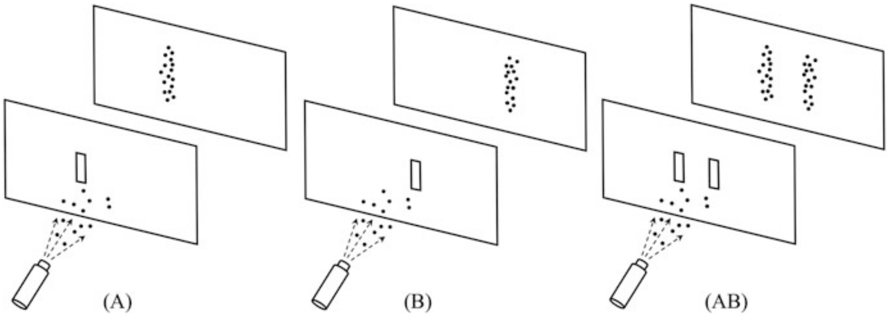


Fig. 2 A schematic description of the classical double-slit experiment, when: (A) only the left slit is open; (B) only the right slit is open; and (AB) both slits are simultaneously open. Note that the time during which the machine gun fired the bullets in situation (AB) is twice than in situations (A) and (B)

are closed. Then, there are no impacts on the detection screen, as no bullets can pass through the barrier. On the other hand, impacts on the detection screen will be observed if (A) the left slit is open and the right one is closed; (B) the right slit is open and the left one is closed; (AB) both slits are open. The distribution of impacts observed in these three configurations is schematically depicted in Fig. 2. As one would expect, the 'both slits open' situation can be easily deduced from the two 'only one-slit open' situations, in the sense that if $\mu_A(x)$ and $\mu_B(x)$ are the probabilities of having an impact at location x on the detection screen, when only the left (resp., the right) slit is open, then the probability $\mu_{AB}(x)$ of having an impact at that same location x, when both slits are kept open, is simply given by the uniform average:

$$\mu_{AB}^{\text{bull}}(x) = \frac{1}{2}[\mu_A(x) + \mu_B(x)]. \tag{1}$$

Consider now a similar experiment, using electrons instead of small bullets. As well as for the bullets, well-localized traces of impact are observed on the detection screen in the situations when only one slit is open at a time, always with the traces of impact distributed in positions that are in proximity of the open slit. On the other hand, as schematically depicted in Fig. 3, when both slits are jointly open, what is obtained is not anymore deducible from the two 'only one-slit open' situations. More precisely, when bullets are replaced by electrons, (1) is not anymore valid and we have instead:

$$\mu_{AB}^{\text{elec}}(x) = \frac{1}{2}[\mu_A(x) + \mu_B(x)] + \text{Int}_{AB}(x), \tag{2}$$

where $\text{Int}_{AB}(x)$ is a so-called *interference contribution*, which corrects the classical uniform average (1) and can take both positive and negative values. Clearly, a corpuscular interpretation of the experiment becomes now impossible, as the region

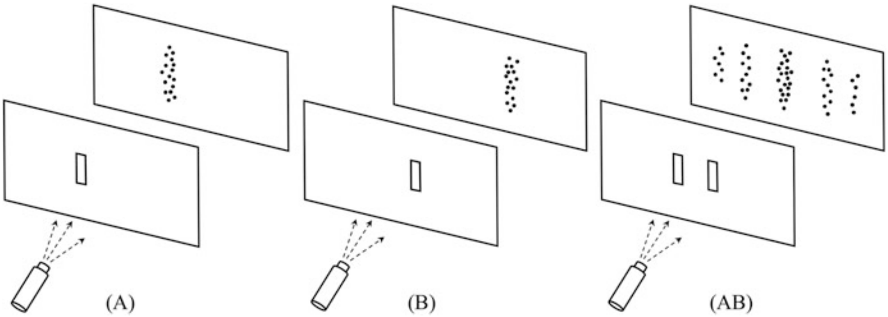


Fig. 3 A schematic description of the quantum double-slit experiment, when: (A) only the left slit is open; (B) only the right slit is open; and (AB) both slits are simultaneously open. Different from the classical (corpuscular) situation, a fringe (interference) pattern appears when the left and right slits are both open

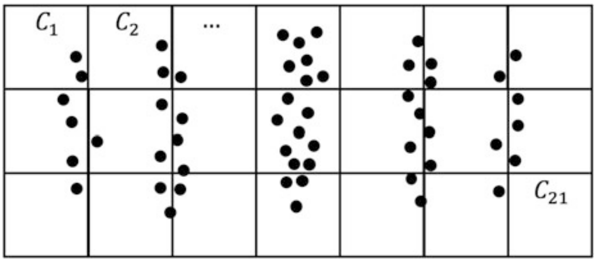


Fig. 4 The detection screen, partitioned into $n = 21$ different cells, each one playing the role of an individual position detector, here showing the traces of $m = 54$ impacts. The experimental probabilities are: $\mu_{AB}(C_1; 21) = \frac{2}{54}$, $\mu_{AB}(C_2; 21) = \frac{2}{54}$, $\mu_{AB}(C_3; 21) = \frac{1}{54}$, $\mu_{AB}(C_4; 21) = \frac{7}{54}, \ldots, \mu_{AB}(C_{20}; 21) = \frac{1}{54}$, $\mu_{AB}(C_{21}; 21) = 0$

where most of the traces of impact are observed is exactly in between the two slits, where instead we would expect to have almost no impacts. Also, in the regions in front of the two slits, where we would expect to have the majority of impacts, practically no traces of impact are observed.

Imagine for a moment that we are only interested in modeling the data of the experiment (either with bullets or electrons) in a very instrumentalistic way, by limiting the description only to what can be observed at the level of the detection screen, i.e., the traces that are left on it. For this, one can proceed as follows. The surface of the detection screen is first partitioned into a given number n of numbered cells $C_1, \ldots, C_n$ (see Fig. 4). Then, the experiment is run until m traces are obtained on it, m being typically a large number. Also, the number of traces of impact in each cell is counted. If $m_{AB}(C_i)$ is the number of traces counted in cell C_i, $i = 1, \ldots, n$, the experimental probability of having an impact in that cell is given by the ratio $\mu_{AB}(C_i; m) = \frac{m_{AB}(C_i)}{m}$. Here by 'experimental probability' we simply mean the probability "induced" by a relative frequency over a large number of repetitions of a same measurement, under the same experimental conditions. Similarly, we have

$\mu_A(C_i; m) = \frac{m_A(C_i)}{m}$ and $\mu_B(C_i; m) = \frac{m_B(C_i)}{m}$, where $m_A(C_i)$ and $m_B(C_i)$ are the number of traces counted in cell C_i when only the left and right slits are kept open, respectively. If the experiments are performed using small bullets, one finds that the difference $\mu_{AB}(C_i; m) - \frac{1}{2}[\mu_A(C_i; m) + \mu_B(C_i; m)]$ tends to zero, as m tends to infinity, for all $i = 1, \ldots, n$, whereas if the experiment is done using micro-entities, like electrons, it does not converge to zero, but towards a function $\text{Int}(C_i)$, expressing the amount of deviation from the uniform average situation.

Now, once the three real functions $\mu_A(C_i; m)$, $\mu_B(C_i; m)$, and $\mu_{AB}(C_i; m)$ have been obtained, and their $m \to \infty$ limit deduced, one could say to have successfully modeled the experimental data, in the three different configurations of the barrier. However, a physicist would not be satisfied with such a modeling. Why? Well, because it is not able to explain why $\mu_{AB}(C_i) = \lim_{m\to\infty} \mu_{AB}(C_i; m)$ cannot be deduced, as one would expect, from $\mu_A(C_i) = \lim_{m\to\infty} \mu_A(C_i; m)$ and $\mu_B(C_i) = \lim_{m\to\infty} \mu_B(C_i; m)$, and why $\mu_{AB}(C_i)$ possesses such a particular interference-like fringe structure. So, let us explain how the quantum explanation typically goes. For this, we will need to exit the two-dimensional plane of the detection screen and describe things at a much more abstract and fundamental level of our physical reality.

As is well-known, even if our description extends from the two-dimensional plane of the detection screen to the three-dimensional theater containing the entire experimental apparatus, this will still be insufficient to explain how the interference pattern is obtained. Indeed, electrons cannot be modeled as spatial waves, as they leave well-localized traces of impact on a detection screen, and they cannot be modeled as particles, as they cannot be consistently associated with trajectories in space.[3] They are truly "something else," which needs to be addressed in more abstract terms. And this is precisely what the quantum formalism is able to do, when describing physical entities in terms of the abstract notions of *states*, *evolutions*, *measurements*, *properties*, and *probabilities*, not necessarily attributable to a description of a spatial (or spatiotemporal) kind.

So, let $|\psi\rangle$ be the state of an electron[4] (at a given moment in time) after having interacted with the double-slit barrier, with both slits open (we use here Dirac's notation). We can consider that this vector state has two components: one corresponding to the electron being reflected back towards the source (assuming for simplicity that the barrier cannot absorb it), and the other one corresponding to the electron having successfully passed through the barrier and reached the detection screen. Let then P_C be the projection operator associated with the property of "having been reflected back by the barrier," and P_{AB} the projection operator associated with the property of "having passed through the two slits." For instance,

[3]This statement remains correct even in the de Broglie–Bohm interpretation of quantum mechanics, as in the latter the trajectories of the micro-quantum entities can only be defined at the price of introducing an additional non-spatial field, called the quantum potential.

[4]One should say, more precisely, that $|\psi_{AB}\rangle$ is a Hilbert-space vector representation of the electron state, as a same state can admit different representations, depending on the adopted mathematical formalism.

P_C could be chosen to be the projection onto the set of states localized in the half-space defined by the barrier and containing the source, whereas P_{AB} would project onto the set of states localized in the other half-space, containing the detection screen.[5] We thus have $P_C + P_{AB} = \mathbb{I}$, and we can define $|\psi_{AB}\rangle = \frac{P_{AB}|\psi\rangle}{\|P_{AB}|\psi\rangle\|}$, which is the state the electron is in after having passed through the barrier and reached the detection screen region. Note that the barrier acts as a filter, in the sense that if the electron does leave a trace on the detection screen, we know it did successfully pass through the barrier, and therefore was in state $|\psi_{AB}\rangle$ when detected.

Now, since by assumption the n cells C_i of the detection screen work as distinct measuring apparatuses, and an electron cannot be simultaneously detected by two different cells, for all practical purposes we can associate them with n orthonormal vectors $|e_i\rangle$, $\langle e_i|e_j\rangle = \delta_{ij}$, corresponding to the different possible outcome-states of the position measurement performed by the screen. This means that we can consider $\{|e_1\rangle, \ldots, |e_n\rangle\}$ to form a basis of the subspace of states having passed through the barrier, and since we are not interested in electrons not reaching the detection screen, we can consider such n-dimensional subspace to be the effective Hilbert space $\mathcal{H}$ of our quantum system, which, for instance, can be taken to be isomorphic to the vector space $\mathbb{C}^n$ of all n-tuples of complex numbers.

According to the Born rule (which in quantum mechanics is used to obtain a correspondence between what is observed in measurement situations, in terms of relative frequencies, and the objects of its mathematical formalism, thus expressing the statistical content of the theory and allowing to bring the latter in contact with the experiments), the probability for an electron in state $|\psi_{AB}\rangle \in \mathcal{H}$, to be detected by cell C_i, is given by the square modulus of the amplitude $\langle e_i|\psi_{AB}\rangle$, that is: $\mu_{AB}(\mathrm{C}_i) = |\langle e_i|\psi_{AB}\rangle|^2$, and if we assume that an electron that has passed through the barrier is necessarily absorbed by the screen (assuming, for instance, that the latter is large enough), we have $\sum_{i=1}^n \mu_{AB}(\mathrm{C}_i) = 1$. Introducing the orthogonal projection operators $P_i = |e_i\rangle\langle e_i|$, we can also write, equivalently:

$$\mu_{AB}(\mathrm{C}_i) = \|P_i|\psi_{AB}\rangle\|^2 = \langle\psi_{AB}|P_i^\dagger P_i|\psi_{AB}\rangle = \langle\psi_{AB}|P_i^2|\psi_{AB}\rangle = \langle\psi_{AB}|P_i|\psi_{AB}\rangle. \tag{3}$$

More generally, if I is a given subset of $\{1, \ldots, n\}$, we can define the projection operator $M = \sum_{i\in I} P_i$, onto the set of states localized in the subset of cells with indexes in I, and the probability of being detected in one of these cells is given by:

$$\mu_{AB}(i \in I) = \langle\psi_{AB}|M|\psi_{AB}\rangle = \sum_{i\in I} \mu_{AB}(\mathrm{C}_i). \tag{4}$$

As an example, consider the situation of Fig. 4, where one can, for instance, define the following seven projectors $M_k = P_k + P_{k+7} + P_{k+14}$, $k = 1, \ldots, 7$,

[5]Intuitively, one can also think of P_{AB} as the projection operator onto the set of states having their momentum oriented towards the detection screen. Of course, all these definitions are only meaningful if applied to asymptotic states, viewing the interaction of the electron with the barrier as a scattering process, with the barrier playing the role of the local scattering potential.

describing the seven columns of the 3 × 7 screen grid. In particular, we have: $\mu_{AB}(i \in \{4, 11, 18\}) = \frac{7}{54} + \frac{8}{54} + \frac{3}{54} = \frac{1}{3}$, i.e., the probability for a trace of impact to appear in the central vertical sector of the screen (the central fringe) is one-third.

The double-slit experiment does not allow to determine if an electron that leaves a trace of impact on the detection screen has passed through the left slit or the right slit. This means that the properties "passing through the left slit" and "passing through the right slit" remain potential properties during the experiment, i.e., alternatives that are not resolved and therefore (as we are going to see) can give rise to interference effects [24]. Let however write P_{AB} as the sum of two projectors: $P_{AB} = P_A + P_B$, where P_A corresponds to the property of "passing through the left slit" and P_B to the property of "passing through the right slit." Note that there is no unique way to define these properties, and the associated projections, as is clear that electrons are not corpuscles moving along spatial trajectories. A possibility here is to further partition the half-space defined by P_{AB} into two sub-half-spaces, one incorporating the left slit, defined by P_A and the other one incorporating the right slit, defined by P_B, so that $P_A P_B = P_B P_A = 0$. For symmetry reasons, we can assume that the electron has no preferences regarding passing through the left or right slits (this will be the case if the source is placed symmetrically with respect to the two slits), so that $\| P_A|\psi_{AB}\rangle\|^2 = \| P_B|\psi_{AB}\rangle\|^2 = \frac{1}{2}$. We can thus define the two orthogonal states $|\psi_A\rangle = \sqrt{2}\, P_A|\psi_{AB}\rangle$ and $|\psi_B\rangle = \sqrt{2}\, P_B|\psi_{AB}\rangle$, and write:

$$|\psi_{AB}\rangle = (P_A + P_B)|\psi_{AB}\rangle = \frac{1}{\sqrt{2}}(|\psi_A\rangle + |\psi_B\rangle). \tag{5}$$

According to the above definitions, $|\psi_A\rangle$ and $|\psi_B\rangle$ can be interpreted as the states describing an electron passing through the left and right slit, respectively.[6] In other words, in accordance with the quantum mechanical *superposition principle*, we have expressed the electron state in the double-slit situation as a (uniform) superposition of one-slit states. Inserting (5) in (4), now omitting the argument in the brackets to simplify the notation, we thus obtain:

$$\begin{aligned}
\mu_{AB} &= \langle\psi_{AB}|M|\psi_{AB}\rangle = \frac{1}{2}(\langle\psi_A| + \langle\psi_B|)M(|\psi_A\rangle + |\psi_B\rangle) \\
&= \frac{1}{2}(\langle\psi_A|M|\psi_A\rangle + \langle\psi_B|M|\psi_B\rangle + \langle\psi_A|M|\psi_B\rangle + \langle\psi_B|M|\psi_A\rangle) \\
&= \frac{1}{2}(\mu_A + \mu_B) + \underbrace{\Re\,\langle\psi_A|M|\psi_B\rangle}_{\mathrm{Int}_{AB}},
\end{aligned} \tag{6}$$

[6]Note however that, as we mentioned already, it is not possible to unambiguously define the two projection operators P_A and P_B, for instance, because of the well-known phenomenon of the spreading of the wave-packet. In other words, there are different ways to decompose $|\psi_{AB}\rangle$ as the superposition of two states that can be conventionally associated with the one-slit situations, as per (5).

where Int_{AB} is the interference contribution, with the symbol $\Re$ denoting the real part of a complex number, and we have used $\langle\psi_B|M|\psi_A\rangle = \langle\psi_A|M|\psi_B\rangle^*$. So, when there are indistinguishable alternatives in an experiment, as is the case here, since we can only observe the traces of the impact in the detection screen, without being able to tell through which slit the electrons have passed, states are typically expressed as a superposition of the states describing these alternatives, and because of that a deviation from the classical probabilistic average (1) will be observed, explaining in particular why an interference-like fringe-like pattern can form.[7]

3 Interrogative Processes

We now want to provide a cognitivistic/conceptualistic interpretation of the double-slit experiment, describing it as an interrogative process [11, 14]. It is of course well understood that measurements in physics' laboratories are like interrogations. Indeed, when we want to measure a physical observable on a given physical entity, we can always say that we have a question in mind, that is: "What is the value of such physical observable for the entity?" By performing the corresponding measurement, we then obtain an answer to the question. More precisely, the outcome of the measurement becomes an input for our human mind, which attaches to it a specific meaning, and it is only when such mental process has been completed that we can say to have obtained an answer to the question that motivated the measurement. In other words, there is a cognitive process, performed by our human mind, and there is a physical process, which provides an input for it.

All this is clear, however, we want to push things further and consider that a measurement can also be described, per se, as an interrogative process, independently of a human mind possibly taking knowledge of its outcome. In other words, we also consider the physical apparatus as a cognitive entity, which answers a question each time it interacts with a physical entity subjected to a measurement, here viewed as a conceptual entity carrying some kind of meaning. This means that two cognitive processes are typically involved in a measurement, one at the level of the apparatus, and another one at the level of the mind of the scientist interacting with it. The latter is founded on human meaning, but not the former, which is the reason why we have to make as humans a considerable effort to understand what is going on. In that respect, we can say that the construction of the theoretical and conceptual edifice of quantum mechanics has been precisely our effort in the attempt to understand the non-human meaning that is exchanged in physical processes, for instance, when an electron interacts with a detection screen in a double-slit experiment.

We will not enter here into the details of this *conceptuality interpretation* of quantum mechanics, and simply refer to the review article [14] and to the references

[7] Of course, to characterize in detail such pattern one should explicitly solve the Schrödinger equation, which however would go beyond the scope of the present text.

cited therein; this not only for understanding the genesis of this interpretation, but also for appreciating why it possibly provides a deep insight into the nature of our physical word. In the following, we limit ourselves to describing the double-slit experiment in a cognitivistic way, as this will be useful when we transpose the approach to an IR-like ambit. So, we start from the hypothesis that the electrons emitted by the electron gun are ‘meaning entities’, i.e., entities behaving in a way that is similar to how human concepts behave. And we also consider the detection screen to be a ‘cognitive entity’, i.e., an entity sensitive to the meaning carried by the electrons and able to answer questions by means of a written (pointillistic) language of traces of impact on its surface. We are then challenged as humans to understand the meaning of this language, and more precisely to guess the query that is answered each time, and then see if the collection of obtained answers is consistent with the logic of such query.

There are of course different equivalent ways to formulate the question answered by the screen detector’s mind. A possible formulation of it is the following: “What is a good example of a trace of impact left by an electron passing through the left slit or the right slit?” This way of conceptualizing the question is of course very “human,” being based on the prejudice that the electron would be an entity always having spatial properties, which is not the case (this depends on its state). But we can here understand the “passing through” concept as a way to express the fact that the probability of detecting the electron by the final screen is zero if both slits are closed. An alternative way of formulating the same question, avoiding the “passing through” concept could be: “What is a good example of an effect produced by an electron interacting with the barrier having both the left and right slits open?” However, we will use in our reasoning the previous formulation of the question, as more intuitive for our spatially biased human minds. What we want is to explain the emergence of the fringe pattern by understanding the process operated by the detection screen, when viewed as a cognitive entity answering the above question.

The first thing to observe is that such process will be generally indeterministic. Indeed, when we say “passing through a slit,” this is not sufficient to specify a unique trajectory in space for an electron (when assumed to be like a spatial corpuscle). This means that, if the screen cognitive entity thinks of the electron as a corpuscle, there are many ways in which it can pass through a slit, so, it will have to select one among several possibilities, which is the reason why, every time the question is asked, the answer (the trace of the impact on the screen) can be different (and cannot be predicted in advance), even though the state of the electron is always the same. The same unpredictability will manifest if the screen cognitive entity does not think of the electron as a spatial entity, but as a more abstract (non-spatial) conceptual entity, which can only acquire spatial properties by interacting with it. Indeed, also in this case the actualization of spatial properties will be akin to a *symmetry breaking* process, whose outcomes cannot be predicted in advance.

To understand how the cognitive process of the screen detector entity might work, let us first concentrate on the central fringe, which is the one exhibiting the higher density of traces of impact and which is located exactly in between the two slits. It is there that the “screen mind” is most likely to manifest an answer. To understand the reason of that, we observe that an impact in that region elicits a maximum doubt as regard the slit the electron would have taken to cross the barrier, or even that it would have necessarily passed through either the left or the right slit, in an exclusive manner. Thus, an impact in that region is a perfect exemplification of the concept "an electron passing through the left slit or the right slit." Now, the two regions on the screen that are exactly opposite the two slits, they have instead a very low density of traces of impact, and again this can be understood by observing that an answer in the form of a trace of impact there would be a very bad exemplification of the concept "an electron passing through the left slit or the right slit," as it would not make us doubt much about the slit taken by the electron. Moving from these two low-density regions, we will then be back in situations of doubt, although less perfect than that of the central fringe, so we will find again a density of traces of impact, but this time less important, and then again regions of low density will appear, and so on, explaining in this way the alternating fringe pattern observed in experiments [11, 14].

4 Modeling the QWeb

Having analyzed the double-slit experiment, and its possible cognitivistic/conceptualistic interpretation, we are now ready to transpose its narrative to the modeling of the meaning entity associated with the Web, which we have called the QWeb. Our aim is to provide a rationale for capturing the full meaning content of a collection of documental entities, which in our case will be the webpages forming the Web, but of course all we are going to say also works for other corpora of documents. As we explained in Sect. 1, there is a universal line for going from abstract concepts to more concrete ones, which is the one going from concepts indicated by single words (or few words) to those that are complex combinations of large numbers of concepts, which in our spatiotemporal theater can manifest as full-fledged stories, and which in our case we are going to associate to the different pages of the Web. Assuming they would have been numbered, we denote them W_i, $i = 1, \ldots, n$. The meaning content of the Web has of course been created by us humans, and each time we interact with the webpages, for instance, when reading them, cognitive processes will be involved, which in turn can give rise to the creation of new webpages. However, we will not be interested here in the modeling of these human cognitive activities, as well as when we model an experiment conducted in a physics' laboratory we are generally not interested in also modeling the cognitive activity of the involved scientists.

As mentioned in Sect. 1, we want to fully exploit the analogy between an IR process, viewed as an interrogation producing a webpage as an outcome, and a measurement, like the position measurement produced by the screen detector in a double-slit experiment, also viewed as being the result of an interrogative process. So, instead of the n cells C_i, $i = 1, \ldots, n$, partitioning the surface of the detection screen, we now have the n webpages W_i, $i = 1, \ldots, n$, partitioning the Web canvas. What we now measure is not an electron, but the QWeb meaning entity, which similarly to an electron we assume can be in different states and can produce different possible outcomes when submitted to measurements. We will limit ourselves to measurements having the webpages W_i as their outcomes. More precisely, webpages W_i will play the same role as the cells C_i of the detection screen in the double-slit experiment, in the sense that we do not distinguish in our measurements the internal structure of a webpage, in the same way that we do not distinguish the locations of the impacts inside a single cell. So, similarly to what we did in Sect. 2, we can associate each webpage with a state $|e_i\rangle$, $i = 1, \ldots, n$, so that $\{|e_1\rangle, \ldots, |e_n\rangle\}$ will form a basis of the n-dimensional QWeb's Hilbert state space.

Let us describe the kind of measurements we have in mind for the QWeb. We will call them "tell a story measurements," and they consist in having the QWeb, prepared in a given state, interacting with an entity sensitive to its meaning, having the n webpages stored in its memory, as stories, so that one of these Web's stories will be told at each run of these measurements, with a probability that depends on the QWeb's state. The typical example of this is that of a search engine having the n webpages stored in its indexes, used to retrieve some meaningful information, with the QWeb initial state being an expression of the meaning contained in the retrieval query (here assuming that the search engine in question would be advanced enough to also use indeterministic processes, when delivering its outcomes).

If the state of the QWeb is $|e_i\rangle$, associated with the webpage W_i, then the 'tell a story measurement' will by definition provide the latter as an outcome, with probability equal to one. But the states $|e_i\rangle$, associated with the stories written in the webpages W_i, only correspond, as we said, to the more concrete states of the QWeb, according to the definition of concreteness given in Sect. 1, and therefore only represent the tip of the iceberg of the QWeb's state space, as it would be the case for the position states of an electron. Indeed, the QWeb's states, in general, can be written as a superposition of the webpages' basis states:

$$|\psi\rangle = \sum_{j=1}^{n} r_j e^{i\rho_j} |e_i\rangle, \quad r_j, \rho_j \in \mathbb{R}, \quad r_j \geq 0, \quad \sum_{j=1}^{n} r_j^2 = 1. \tag{7}$$

We can right away point out an important difference between (7) and what is usually done in IR approaches, like the so-called *vector space models* (VSM), where the states that are generally written as a superposition of basis states are those associated with the index terms used in queries (see, for instance, [30, p. 5], and [27, p. 19]). Here it is exactly the other way around: the dimension of the state space is determined by the number of available documents, associated with the outcome-states of the 'tell a story measurements', interpreted as stories, i.e., as the more concrete states of the QWeb entity subjected to measurements. This also means that (as we will explain in the following) the states associated with single terms will not necessarily be mutually orthogonal, i.e., will not generally form a basis. Of course, another important difference with respect to traditional IR approaches is that the latter are built upon real vector spaces, whereas our quantum modeling is intrinsically built upon complex vector spaces (Hilbert spaces), where linearity works directly at the level of the complex numbers and weights are only obtained from the square of their moduli. In other words, the complex numbers $r_j e^{i\rho_j}$, appearing in the expansion (7), can be understood as generalized coefficients expressing a connection between the meaning carried by the QWeb in state $|\psi\rangle$, and the meaning "sticking out" from (the stories contained in) the webpages W_j.[8]

As a very simple example of initial state, we can consider a state $|\chi\rangle$ expressing a *uniform meaning connection* towards all the Web stories: $|\chi\rangle = \frac{1}{\sqrt{n}} \sum_{j=1}^n e^{i\rho_j} |e_j\rangle$, so that the probability to obtain story W_i, in a 'tell a story measurement', when the QWeb is in such uniform state $|\chi\rangle$, is:

$$\mu(\mathrm{W}_i) = \langle\chi|P_i|\chi\rangle = \frac{1}{n} \sum_{j,k=1}^{n} e^{i(\rho_j - \rho_k)} \underbrace{\langle e_k|e_i\rangle}_{\delta_{ki}} \underbrace{\langle e_i|e_j\rangle}_{\delta_{ij}} = \frac{1}{n}. \tag{8}$$

As another simple example, we can consider the QWeb state $|\chi_I\rangle = \frac{1}{\sqrt{m}} \sum_{j\in I} e^{i\rho_j} |e_j\rangle$, which is uniform only locally, i.e., such that only a subset I of m webpages , with $m \leq n$, would have the same (non-zero) probability of being selected as an actual story, so that in this case $\mu_I(\mathrm{W}_i) = \langle\chi_I|P_i|\chi_I\rangle = \frac{1}{m}$, if $i \in I$, and zero otherwise.

It is important to observe that we are here viewing the QWeb as a whole entity, when we speak of its states, although it is clearly also a composite entity, in the sense that it is a complex formed by the combination of multiple concepts. Take two concepts A and B (for example, $A =$ *Fruits* and $B =$ *Vegetables*). As individual conceptual entities, they are certainly part of the QWeb composite entity, and as such they can also be in different states, which we can also write as linear combinations of the webpages' basis states:

$$|\psi_A\rangle = \sum_{j=1}^{n} a_j e^{i\alpha_j} |e_j\rangle, \quad |\psi_B\rangle = \sum_{j=1}^{n} b_j e^{i\beta_j} |e_j\rangle, \tag{9}$$

with $a_j, b_j, \alpha_j, \beta_j \in \mathbb{R}$, $a_j, b_j \geq 0$, and $\sum_{j=1}^n a_j^2 = \sum_{j=1}^n b_j^2 = 1$. These states, however, will be considered to be also states of the QWeb entity as a whole, as they also belong to its n-dimensional Hilbert space. In other words, even if states are all considered to be here states of the QWeb entity, some of them will also be interpreted as describing more specific individual conceptual entities forming the QWeb. We thus consider that individual concepts forming the composite QWeb entity can be viewed as specific states of the latter. Of course, the quantum formalism also offers another way to model composite entities, by taking the tensor product of the Hilbert spaces of the sub-entities in question. This is also a possibility, when

[8]More precisely, the real positive number r_j can receive a specific interpretation as quantum *meaning bonds*; see Appendix 2.

modeling conceptual combinations, which proved to be very useful in the quantum modeling of data from cognitive experiments, particularly in relation to the notion of entanglement (see [6, 7] and the references cited therein), but in the present analysis we focus more directly on the superposition principle (and the interference effects it subtends) as a mechanism for accounting for the emergence of meaning when concepts are considered in a combined way [2] (see however the discussion in the first part of Sect. 5).

Since we are placing ourselves in the same paradigmatic situation of the double-slit experiment, we want to consider how the combination of two concepts A and B—let us denote the combination AB—can manifest at the level of the Web stories, in the ambit of a "tell a story measurement." Here we consider the notion of "combination of two concepts" in a very general way, in the sense that we do not specify how the combination of A and B is actually implemented, at the conceptual level. In human language, if A is the concept *Fruits* and B is the concept *Vegetables*, their combination can, for instance, be *Fruits–vegetables*, *Fruits and vegetables*, *Fruits or vegetables*, *Fruits with vegetables*, *Fruits are sweeter than vegetables*, etc., which of course carry different meanings, i.e., describe different states of their two-concept combination. In fact, also stories which are jointly about *Fruits* and *Vegetables* can be considered to be possible states of the combination of these two concepts. All these possibilities give rise of different states $|\psi_{AB}\rangle$, describing the combination of the two concepts A and B.

These two concepts can be seen to play the same role of the two slits in the double-slit experiment. When the two slits are jointly open, we are in the same situation as when the two concepts A and B are jointly considered in the combination AB, producing a state $|\psi_{AB}\rangle$ that we can describe as the superposition of two states $|\psi_A\rangle$ and $|\psi_B\rangle$, which are the states of the concepts A and B, respectively, when considered not in a combination, and which play the same role as the states of the electron in the double-slit experiment traversing the barrier when only one of the two slits is kept open at a time. Of course, different superposition states can in principle be defined, each one describing a different state of the combination of the two concepts, but here we limit ourselves to the superposition (5), where the states $|\psi_A\rangle$ and $|\psi_B\rangle$ have the exact same weight in the superposition.

Let now X be a given concept. It can be a concept described by a single word or a more complex concept described by the combination of multiple concepts. We consider the projection operator M_X^w, onto the set of states that are *manifest* stories about X. This means that we can write:

$$M_X^w = \sum_{i \in J_X} |e_i\rangle\langle e_i|, \tag{10}$$

where J_X is the set of indexes associated with the webpages that are manifest stories about X, where by "manifest" we mean stories that explicitly contain the word(s) "X" indicating the concept X, hence the superscript "w" in the notation, which stands for "word." Indeed, we could as well have defined a more general projection operator $M_X^s = \sum_{i \in I_X} |e_i\rangle\langle e_i|$, onto the set of states that are stories about X not necessarily of the manifest kind, i.e., not necessarily containing the explicit word(s) indicating the concept(s) the stories are about, with $J_X \subset I_X$, and the superscript "s" now standing for "story."

To avoid possible confusions, we emphasize again the difference between the notion of *state of a concept* and that of *story about a concept*. The latter, in our definition, is a webpage, i.e., a full-fledged printed or printable document. But webpages that are stories about a concept may explicitly contain the word indicating such concept or not. For example, one can conceive a text explaining what *Fruits* are, without ever writing the word "fruits" (using in replacement other terms, like "foods in the same category of pineapple, pears, and bananas"). On the other hand, the notion of state of a concept expresses a condition which cannot in general be reduced to that of a story, as it can also be a superposition of stories of that concept (or better, a superposition of the states associated with the stories of that concept), as expressed, for instance, in (7) and (9), and a superposition of (states of) stories is not anymore a (state of a) story.

Now, when considering a "tell a story measurement," we can also decide to only focus on stories having a predetermined content. In the double-slit experiment, this would correspond to only be interested in the detection of the electron by a certain subset of cells, indicated by a given set of indexes J_X, and not the others. More specifically, we can consider only those stories that are "stories about X," where X is a given concept. This means that if the QWeb is in a pre-measurement state $|\psi_A\rangle$, which is the state of a given concept A, what we are asking through the measurement is if the stories about X are good representatives of A in state $|\psi_A\rangle$ (in the same way we can ask if a certain subset of traces of impact, say those of the central fringe, is a good example of electrons passing through the left slit; see the discussion of Sect. 3). In other words, we are asking how much $|\psi_A\rangle$ is meaning connected to concept X, when the latter is in one of the maximally concrete states defined by the webpages that are "stories of X" or even more specifically "manifest stories of X."

In the latter case, we can test this by using the projection operator M_X^w and the Born rule. According to (4), the probability μ_A with which the concept A in state $|\psi_A\rangle$ is evaluated to be well represented by a "manifest story about X" is given by the average:

$$\mu_A(i \in J_X) = \langle\psi_A|M_X^w|\psi_A\rangle = \sum_{i \in J_X} |\langle e_i|\psi_A\rangle|^2 = \sum_{i \in J_X} a_i^2, \tag{11}$$

where for the last equality we have used (9). If we additionally assume that A is more specifically described by a state that is a superposition only of those stories that explicitly contains the words "A" (manifest stories about A), the above probability becomes (omitting from now on the argument, to simplify the notation): $\mu_A = \sum_{i \in J_{A,X}} a_i^2$, where $J_{A,X}$ denotes the sets of indexes associated with the webpages jointly containing the words "A" and "X." Note that if $n_{A,X} = |J_{A,X}|$ is the number of webpages containing both terms "A" and "X," $n_A = |J_A|$ and $n_X = |J_X|$ are the webpages containing the "A" term and the "X" term, respectively, we have $n_{A,X} \leq$

n_A and $n_{A,X} \leq n_X$. Becoming even more specific, we can consider states of A expressing a uniform meaning connection towards all the different manifest stories about A, that is, characteristic function states of the form:

$$|\chi_A\rangle = \frac{1}{\sqrt{n_A}} \sum_{j \in J_A} e^{i\alpha_j} |e_j\rangle, \tag{12}$$

for which the probability (11) becomes:

$$\mu_A = \langle \chi_A | M_X^w | \chi_A \rangle = \sum_{i \in J_{A,X}} \frac{1}{n_A} = \frac{n_{A,X}}{n_A}, \tag{13}$$

which can be simply interpreted as the probability of randomly selecting a webpage containing the term "X," among those containing the terms "A."

With respect to the double-slit experiment analogy, the probability μ_A describes the "only left slit open" situation, and of course, *mutatis mutandis*, we can write (with obvious notation) an equivalent expression for a different concept B: $\mu_B = \langle \chi_B | M_X^w | \chi_B \rangle = \sum_{i \in J_{B,X}} \frac{1}{n_B} = \frac{n_{B,X}}{n_B}$. So, when calculating the probability μ_{AB} for the combination AB of the two concepts A and B, we are in a situation equivalent to when the two slits are kept jointly open, with the question asked being now about the meaning connection between AB, in state $|\psi_{AB}\rangle$, and a (here manifest) story about X. Concerning the state $|\psi_{AB}\rangle$, describing the combination, we want it to be able to account for the emergence of meanings that can possibly arise when the two concepts A and B are considered one in the context of the other, and for consistency reasons we expect the probability μ_{AB} to be equal to $\frac{n_{AB,X}}{n_{AB}}$ (since we are here limiting our discussion, for simplicity, to manifest stories), where n_{AB} is the number of webpages containing both the "A" and "B" terms and $n_{AB,X}$ is the number of webpages containing in addition also the "X" term, and of course: $n_{AB,X} \leq n_{AB}$, $n_{AB} \leq n_A$, and $n_{AB} \leq n_B$. This can be easily achieved if the state of AB is taken to be the characteristic function state: $|\chi_{AB}\rangle = \frac{1}{\sqrt{n_{AB}}} \sum_{j \in J_{AB}} e^{i\delta_j} |e_j\rangle$; however, coming back to our discussion of Sect. 2, this would not be a satisfactory way to proceed, as the modeling would then remain at the level of the canvas of printed documents of the Web, and would therefore not be able to capture the level of meaning associated with it, that is, the more abstract QWeb entity. It is only at the level of the latter that emergent meanings can be explained as the result of combining concepts.

By analogy with the paradigmatic double-slit experiment, we will here assume that a state of AB, i.e., a state of the combination of the two concepts A and B, when they are in individual states $|\psi_A\rangle$ and $|\psi_B\rangle$, respectively, can be generally represented as a superposition vector (5). Since here we are considering the special case where these states are characteristic functions, we more specifically have:

$$|\psi_{AB}\rangle = \frac{1}{\sqrt{2}}(|\chi_A\rangle + |\chi_B\rangle), \tag{14}$$

where we have assumed for simplicity that $|\chi_A\rangle$ and $|\chi_B\rangle$ can be taken to be orthogonal states (this need not to be the case in general). The interference contribution $\text{Int}_{AB} = \Re\,\langle\chi_A|M_X^w|\chi_B\rangle$ can then be calculated by observing that:

$$M_X^w|\chi_B\rangle = \left(\sum_{j\in J_X}|e_j\rangle\langle e_j|\right)\left(\frac{1}{\sqrt{n_B}}\sum_{k\in J_B}e^{i\beta_k}|e_k\rangle\right)$$
$$= \frac{1}{\sqrt{n_B}}\sum_{j\in J_X}\sum_{k\in J_B}e^{i\beta_k}|e_j\rangle\underbrace{\langle e_j|e_k\rangle}_{\delta_{jk}} = \frac{1}{\sqrt{n_B}}\sum_{j\in J_{B,X}}e^{i\beta_j}|e_j\rangle, \qquad (15)$$

so that, multiplying the above expression from the left by $\langle\chi_A|$ and taking the real part, we obtain:

$$\text{Int}_{AB} = \Re\left(\frac{1}{\sqrt{n_A}}\sum_{j\in J_A}e^{-i\alpha_j}\langle e_j|\right)\left(\frac{1}{\sqrt{n_B}}\sum_{k\in J_{B,X}}e^{i\beta_k}|e_k\rangle\right)$$
$$= \frac{1}{\sqrt{n_A n_B}}\sum_{j\in J_A}\sum_{k\in J_{B,X}}\underbrace{\langle e_j|e_k\rangle}_{\delta_{jk}}\underbrace{\Re\,e^{i(\beta_k-\alpha_j)}}_{\cos(\beta_k-\alpha_j)} = \sum_{j\in J_{AB,X}}\frac{\cos(\beta_j-\alpha_j)}{\sqrt{n_A n_B}}. \qquad (16)$$

According to (6), (13), and (16), the probability μ_{AB} for the combined concept AB is therefore:

$$\mu_{AB} = \frac{1}{2}\Big(\underbrace{\frac{n_{A,X}}{n_A}}_{\mu_A}+\underbrace{\frac{n_{B,X}}{n_B}}_{\mu_B}\Big)+\sum_{j\in J_{AB,X}}\frac{\cos(\beta_j-\alpha_j)}{\sqrt{n_A n_B}}. \qquad (17)$$

It is important to observe in (17) the role played by the phases α_j and β_j characterizing the states $|\chi_A\rangle$ and $|\chi_B\rangle$. When they are varied, the individual probabilities μ_A and μ_B remain perfectly invariant, whereas the values of μ_{AB} can explore an entire range of values, within the interference interval $I_{AB} = [\mu_{AB}^{\min}, \mu_{AB}^{\max}]$, where according to (17) we have:

$$\mu_{AB}^{\min} = \frac{1}{2}\left(\frac{n_{A,X}}{n_A}+\frac{n_{B,X}}{n_B}\right)-\frac{n_{AB,X}}{\sqrt{n_A n_B}},$$
$$\mu_{AB}^{\max} = \frac{1}{2}\left(\frac{n_{A,X}}{n_A}+\frac{n_{B,X}}{n_B}\right)+\frac{n_{AB,X}}{\sqrt{n_A n_B}}. \qquad (18)$$

Therefore, we see that via the interference effects, the co-occurrence of the terms "A," "B," and "X" is independent of what is revealed in the Web for the co-occurrence of just "A" and "X" or the co-occurrence of just "B" and "X." This means that it is really at the more abstract level of the QWeb, and not of the Web, that these three situations of co-occurrence can be seen to be related to each other.

5 Adding Context

According to (18), by using the superposition principle and the corresponding interference effects, we can extend the values of the probability μ_{AB} beyond those specified by the uniform average $\mu_{AB}^{\text{uni}} = \frac{1}{2}\left(\frac{n_{A,X}}{n_A} + \frac{n_{B,X}}{n_B}\right)$. One may wonder then if, generally speaking, interference effects would be sufficient to model all possible situations. The answer is negative, and to see why let us consider a simple example of a collection of documents for which interference effects are insufficient for their modeling.[9]

Assume that the collection is formed by n documents ($n \geq 140$) that $n_A = 100$ of them contain a given word "A," and $n_B = 50$ of them contain another word "B." Also, the number of documents containing both words is assumed to be $n_{AB} = 10$ (see Fig. 5). Consider then a third word "X," which is assumed to be present in 80 of the documents containing the word "A," in 15 of the documents containing the word "B," and in 5 of the documents containing both words, that is: $n_{A,X} = 80$, $n_{B,X} = 15$, $n_{AB,X} = 5$. So, $\mu_A = \frac{n_{A,X}}{n_A} = \frac{80}{100} = 0.8$, $\mu_B = \frac{n_{B,X}}{n_B} = \frac{15}{50} = 0.3$, and $\mu_{AB}^{\text{uni}} = \frac{1.1}{2} = 0.55$. We also have, $\frac{n_{AB,X}}{\sqrt{n_A n_B}} = \frac{5}{\sqrt{5000}} \approx 0.07$, so that $\mu_{AB}^{\min} \approx 0.55 - 0.07 = 0.48$ and $\mu_{AB}^{\max} \approx 0.55 + 0.07 = 0.62$.

Now, as we said, μ_{AB}, for consistency reasons, should be equal to $\frac{n_{AB,X}}{n_{AB}} = \frac{5}{10} = 0.5$, i.e., to the probability of randomly selecting a document containing the word "X," among those containing the words "A" and "B." Since 0.5 is contained in the interference interval $I_{AB} = [0.48, 0.62]$, by a suitable choice the phase differences in (17), the equality $\mu_{AB} = \frac{n_{AB,X}}{n_{AB}}$ can be obtained; hence, interference effects are sufficient to model this situation. But if we consider a word "Y" that, different from "X," would only be present in 10 of the documents containing the word "A" and in 10 of those containing the word "B" (see Fig. 5), this time we have $\mu_A = \frac{n_{A,Y}}{n_A} = \frac{10}{100} = 0.1$, $\mu_B = \frac{n_{B,Y}}{n_B} = \frac{10}{50} = 0.2$, and $\mu_{AB}^{\text{uni}} = \frac{0.3}{2} = 0.15$. So, $\mu^{\min} \approx 0.15 - 0.07 = 0.08$ and $\mu^{\max} \approx 0.15 + 0.07 = 0.22$, which means that $\frac{n_{AB,Y}}{n_{AB}} = 0.5$ is not anymore contained in the interference interval $I_{AB} = [0.08, 0.22]$. Hence, interference effects are not sufficient to model this situation.

Additional mechanisms should therefore be envisioned to account for all the probabilities that can be calculated by counting the relative number of documents containing certain words and co-occurrences of words. A possibility is to explore

[9]The example is taken from [4]. Note however that the two situations described in [4] required both the use of "interference plus context effects," contrary to what was stated in the article. Here we provide a corrected version of the example, where the first situation only requires interference effects, whereas the second situation requires interference plus context effects.

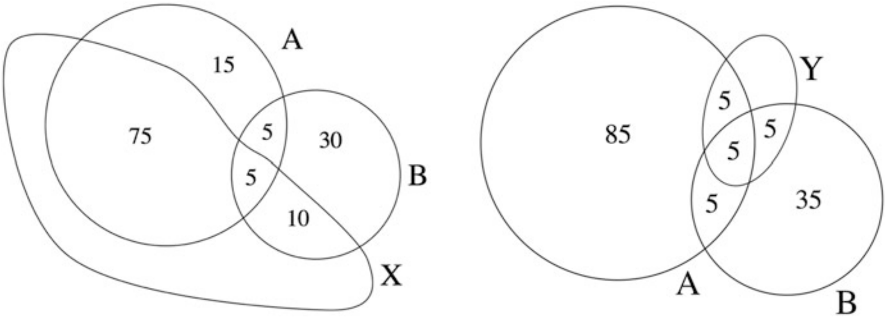


Fig. 5 A schematic Venn-diagram representation of the number of documents containing the words "A," "B," and "X" (left) which can be modeled using only interference effects, and the words "A," "B," and "Y" (right), which instead also require context effects

more general forms of measurements on more general versions of the QWeb entity. In our approach here, we focused on the superposition principle to account for the emergence of new meanings when concepts are combined. But of course, when a cognitive entity interacts with a meaning entity, the emergence of meaning is not the only element that might play a role. In human reasoning, for instance, a two-layer structure can be evidenced: one consisting of *conceptual thoughts*, where a combination of concepts is evaluated as a new single concept, and the other consisting of *classical logical thoughts*, where a combination of concepts is evaluated as a classical combinations of different entities [17].

To also account for the existence of classical logical reasoning, one can define more general "tell a story measurements," by considering a specific type of Hilbert space called *Fock space*, originally used in *quantum field theory* to describe situations where there is a variable number of identical entities. This amounts considering the QWeb as a more general "quantum field entity" that can be in different *number operator* states and in different superpositions of these states. In the present case, since we are only considering the combination of two concepts, the construction of the Fock space $\mathcal{F}$ can be limited to two sectors: $\mathcal{F} = \mathcal{H} \oplus (\mathcal{H} \otimes \mathcal{H})$, where "$\oplus$" denotes a *direct sum* between the first sector $\mathcal{H}$ (isomorphic to $\mathbb{C}^n$) and the second sector $\mathcal{H} \otimes \mathcal{H}$ (isomorphic to $\mathbb{C}^{2n}$), where "$\otimes$" denotes the *tensor product*. The first sector describes the one-entity states, where the combination of the two concepts A and B is evaluated as a new (emergent) concept, typically described by a superposition state (5). The second sector describes the two-entity situation, where the two concepts A and B remain separate in their combination, which is something that can be described by a so-called *product* (non-entangled) state $|\psi_A\rangle \otimes |\psi_B\rangle$.

Instead of (5), we can then consider the more general superposition state:

$$|\psi_{AB}\rangle = \sqrt{1-m^2}\, e^{i\nu} \frac{1}{\sqrt{2}}(|\psi_A\rangle + |\psi_B\rangle) + m\, e^{i\lambda} |\psi_A\rangle \otimes |\psi_B\rangle, \tag{19}$$

where the number $0 \le m \le 1$ determines the degree of participation in the second sector. Also, instead of (10), we have to consider a more general projection operator, acting now on both sectors. Here we can distinguish the two paradigmatic projection operators:

$$M_X^{w,\text{and}} = M_X^w \oplus (M_X^w \otimes M_X^w), \quad M_X^{w,\text{or}} = M_X^w \oplus (M_X^w \otimes \mathbb{I} + \mathbb{I} \otimes M_X^w - M_X^w \otimes M_X^w), \tag{20}$$

where $M_X^{w,\text{and}}$ describes the situation where the combination of concepts is logically evaluated as a *conjunction* (and), whereas $M_X^{w,\text{or}}$ describes the situation where the combination of concepts is logically evaluated as a *disjunction* (or). When we use $M_X^{w,\text{and}}$, one finds, in replacement of (6), the more general formula[10]:

$$\mu_{AB} = m^2 \, \mu_{AB}^{\text{and}} + (1 - m^2) \left[\frac{1}{2}(\mu_A + \mu_B) + \text{Int}_{AB} \right], \tag{21}$$

where $\mu_{AB}^{\text{and}} = \mu_A \mu_B$. However, this will not be sufficient to model all possible data, as is clear that in the previously mentioned example of word "Y," we have: $\mu_{AB}^{\text{and}} = 0.02$, so that the interval of values that can be explored by the above convex combination (by varying not only the phases α_j and β_j, but now also the coefficient m) is [0.02, 0.22], which still doesn't contain the value 0.5 of $\frac{n_{AB,Y}}{n_{AB}}$. When we use instead $M_X^{w,\text{or}}$, we have to replace μ_{AB}^{and} in (21) by $\mu_{AB}^{\text{or}} = \mu_A + \mu_B - \mu_A \mu_B$, whose value for the word "Y" of our example is 0.28, so that the interval of possible values becomes [0.08, 0.28], which however is still not sufficient.

So, we must find some other cognitive effects, in order to be able to model and provide an explanation for a wide spectrum of experimental values for the probabilities, related to different possible collections of documental entities. A general way of proceeding, remaining in a "first sector" modeling of the QWeb, is to consider that there would be also *context effects* that can alter the QWeb state before it is measured. In the double-slit experiment analogy, we can imagine a mask placed somewhere in between the barrier and the screen, acting as a filter allowing certain states to pass through, whereas others will be blocked (see Fig. 6). Note that if we place the mask close to the detection screen, some cells will be deactivated, as the components of the pre-measurement state relative to them will be filtered out by the mask. On the other hand, if it is placed close to the double-slit barrier, it will allow to control the transmission through the slits and produce, by changing its position, a continuum of interference figures, for instance, interpolating the probability distributions of the two one-slit arrangements; see [21]. More complex effects can of course be obtained if the mask is placed at some finite distances from the barrier and screen, and more general filters than just masks can also be considered, but their overall effect will always be that only certain states will be allowed to interact with the measuring apparatus (here the screen).

[10]It is not in the scope of the present chapter to enter into the details of this Fock space modeling and we simply refer the interested reader to [2, 16, 18].

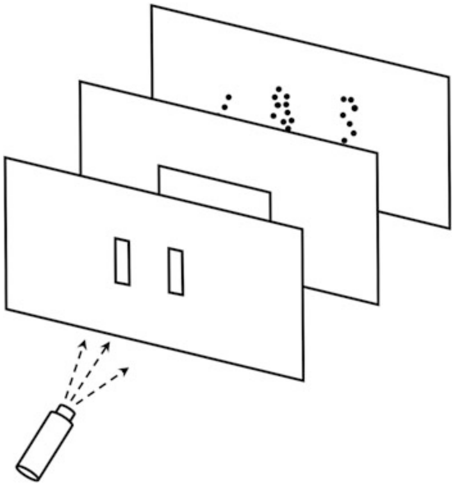


Fig. 6 By placing a screen with a mask (and more generally a filter) between the barrier and the detection screen, the structure of the observed interference pattern can be modulated. The effect of this additional structure can be ideally described using a projection operator

From a cognitivistic viewpoint, context effects can have different origins and logics. For instance, we can consider that an interrogative context, for the very fact that a given question is asked, will inevitably alter the state of the meaning entity under consideration. Even more specifically, consider the example of a cognitive entity that is asked to tell a story (it can be a person, a search engine, or the combination of both). For this, a portion of the entity's memory needs to become accessible, and one can imagine that the extent and nature of such available portion of memory can depend on the story that is being asked.[11]

So, we will now assume that when the QWeb entity is subjected to a "tell a story measurement," there will be a preliminary change of state, and we will adopt the very simple modeling of such state change by means of an orthogonal projection operator, which in general can also depend on the choice of stories we are interested in, like "stories about X," so we will generally write N_X for it ($N_X^2 = N_X = N_X^\dagger$). Just to give a simple example of a X-dependent projection N_X, it could be taken to be the projection operator onto the subspace of QWeb's states that are "states of X" (we recall that a "state of X" is generally not necessarily also a "story about X"). However, in the following we will just limit ourselves to the idealization that context effects can be formally modeled using a projection operator, without specifying their exact nature and origin. So, the presence of this additional context produces the pre-measurement transitions: $|\psi_A\rangle \to |\psi'_A\rangle$, $|\psi_B\rangle \to |\psi'_B\rangle$, and $|\psi_{AB}\rangle \to |\psi'_{AB}\rangle$, where we have defined (from now on, for simplicity, we just write N for N_X, dropping the X-subscript):

[11] In the IR ambit, this can also be associated with constraints related to geographical locations and search histories [26, 27].

$$|\psi'_A\rangle = \frac{N|\psi_A\rangle}{\|N|\psi_A\rangle\|}, \quad |\psi'_B\rangle = \frac{N|\psi_B\rangle}{\|N|\psi_B\rangle\|}, \quad |\psi'_{AB}\rangle = \frac{N|\psi_{AB}\rangle}{\|N|\psi_{AB}\rangle\|}. \tag{22}$$

With the above re-contextualized states, the probability $\mu_A = \langle\psi'_A|M^w_X|\psi'_A\rangle$ becomes:

$$\mu_A = \frac{\langle\psi_A|N^\dagger M^w_X N|\psi_A\rangle}{\|N|\psi_A\rangle\|^2} = \frac{\langle\psi_A|NM^w_X N|\psi_A\rangle}{\langle\psi_A|N|\psi_A\rangle} = \frac{1}{p_A}\langle\psi_A|NM^w N|\psi_A\rangle, \tag{23}$$

where for the second equality we have used $\|N|\psi_A\rangle\|^2 = \langle\psi_A|N^\dagger N|\psi_A\rangle = \langle\psi_A|N^2|\psi_A\rangle = \langle\psi_A|N|\psi_A\rangle$, and for the last equality we have defined the probability $p_A = \langle\psi_A|N|\psi_A\rangle$, for the state $|\psi_A\rangle$ to be an eigenstate of the context N. Similar expressions clearly hold also for the concept B: $\mu_B = \frac{1}{p_B}\langle\psi_B|NM^w N|\psi_B\rangle$, with $p_B = \langle\psi_B|N|\psi_B\rangle$, and for the probability $\mu_{AB} = \langle\psi'_{AB}|M^w_X|\psi'_{AB}\rangle$, relative to the concept combination AB, we now have:

$$\begin{aligned}\mu_{AB} &= \frac{\langle\psi_{AB}|N^\dagger M^w_X N|\psi_{AB}\rangle}{\|N|\psi_{AB}\rangle\|^2} = \frac{\langle\psi_{AB}|NM^w_X N|\psi_{AB}\rangle}{\langle\psi_{AB}|N|\psi_{AB}\rangle} \\ &= \frac{\langle\psi_A|NM^w_X N|\psi_A\rangle + \langle\psi_B|NM^w_X N|\psi_B\rangle + 2\Re\,\langle\psi_A|NM^w_X N|\psi_B\rangle}{\langle\psi_A|N|\psi_A\rangle + \langle\psi_B|N|\psi_B\rangle + 2\Re\,\langle\psi_A|N|\psi_B\rangle} \\ &= \frac{p_A\,\mu_A + p_B\,\mu_B + 2\Re\,\langle\psi_A|NM^w_X N|\psi_B\rangle}{p_A + p_B + 2\Re\,\langle\psi_A|N|\psi_B\rangle}.\end{aligned} \tag{24}$$

The first two terms at the numerator of (24) correspond to a *weighted average*, whereas the third term, both at the numerator and denominator, is the interference-like contribution. Note that in the special case where $|\psi_A\rangle$ and $|\psi_B\rangle$ are eigenstates of the context N, that is, $N|\psi_A\rangle = |\psi_A\rangle$ and $N|\psi_B\rangle = |\psi_B\rangle$, we have $p_A = p_B = 1$, so that (24) reduces to (6), or, if $|\psi_A\rangle$ and $|\psi_B\rangle$ are not orthogonal vectors, to:

$$\mu_{AB} = \frac{\frac{1}{2}(\mu_A + \mu_B) + \Re\,\langle\psi_A|M^w_X|\psi_B\rangle}{1 + \Re\,\langle\psi_A|\psi_B\rangle}, \tag{25}$$

where the weighted average now becomes a uniform one. The more general expression (24), incorporating both context and interference effects, allows to cover a much larger range of values. In fact, as we show in Appendix 1, under certain assumptions the full [0, 1] interval of values can be spanned, thus allowing all possible data about occurrence and co-occurrence of words to be modeled.

6 Conclusion

In this chapter, we have motivated a fundamental distinction between the Web of printed pages (or any other collection of documental entities) and a more abstract entity of meaning associated with it, which we have called the QWeb, for which we have proposed a Hilbertian (Born rule based) quantum model. In our discussion, we have focused on an important class of measurements, which we have called the 'tell a story measurements', whose outcome-states are associated with the n webpages and were taken to form a basis of the (n-dimensional) Hilbert space. We have tested the model by considering the specific situation where only stories manifestly containing the words denoting certain concepts are considered, in order to allow to relate the theoretical probabilities with those obtained by calculating the relative frequency of occurrence and co-occurrence of these words, which in turn depend on how much the associated concepts are meaning-connected. We have done so by also considering context effects, in addition to interference effects, the former being modeled by means of orthogonal projection operators and the latter by means of superposition states. Also, we have extensively used the double-slit experiment as a guideline to motivate the transmigration of fundamental notions from physics to human cognition and theoretical computer science.

Note that more general models than those explored here can also be considered, exploiting more general versions of the quantum formalism, like the GTR-model and the extended Bloch representation of quantum mechanics [8–10, 12, 13]. Hence, the "Q" in "QWeb" refers to a *quantum structure* that need not to be understood in the limited sense of the standard quantum formalism. We have also mentioned in Sect. 5 the possibility of working in a multi-sector Fock space, as a way to extend the range of probabilities that can be modeled. However, we observed that not all values can be modeled in this way. Another direction that can be explored (as an alternative to context effects) is to consider states whose meaning connections are not necessarily uniform, although still localized within the sets J_A and J_B. A further other direction is to consider *step function* states extending beyond the manifest word subspaces. For example, states of the form: $|\psi_A^a\rangle = a\,|\chi_A\rangle + \bar{a}\,|\bar{\chi}_A\rangle$, where $|\bar{\chi}_A\rangle = \frac{1}{\sqrt{n-n_A}}\sum_{j\notin J_A} e^{i\alpha_j}|e_j\rangle$, and $|a|^2 + |\bar{a}|^2 = 1$.

Regarding the co-occurrences of words in documents, it is worth observing that they are determined by the meaning carried by the corresponding concepts and documents, and not by the physical properties of the latter. This means that we can access the traces left by meaning by analyzing the co-occurrence of words in the different physical (printed or stored in memory) documents, and that such meaning "stick out" from the latter in ways that can be accessed without the intervention of the human minds that created it. Note however that the meaning extending out of these documents, here the webpages, is not the full meaning of the QWeb, as encoded in its quantum state. This is so because one cannot reconstruct the pre-measurement state of a quantum measurement by having only access to the outcome (collapsed) states and the associated probabilities of a single measurement. For this, one needs to perform a series of different measurements, characterized by different *informationally complete* bases, as is done in the so-called *quantum state tomography* [22]. Here we only considered the basis associated with the webpages, and it is still unclear which complementary measurements could be defined, using different bases and having a clear operational meaning, that is, which can be concretely performed, at least in principle [4].

Let us also observe that, generally speaking, in IR situations also the modeling of how human minds interact with the QWeb can and will play a role, in addition to the modeling per se of the QWeb. Indeed, as we mentioned already in Sect. 3, the outcome provided by a measurement of the QWeb, say a given story in a "tell a story measurement," becomes the input with which human minds will have to further interact with, which again can be described as a deterministic or indeterministic context, possibly creating new meanings. The formalism of quantum theory can again be used to model these human cognitive interactions, which is what is typically investigated in cognitive psychology experiments and again modeled using the mathematical formalism of quantum theory, in the emerging field known as *quantum cognition*; see [15] and the references cited therein.

We stress that, in our view, it is only when a more abstract—meaning oriented—approach is adopted in relation to documental entities, like the Web, and an operational-realistic modeling of its conceptual structure is attempted, exploiting the panoply of quantum effects that have been discovered in the physics' laboratories, that quoting from [4]: "a deeper understanding of how meaning can leave its traces in documents can be accessed, possibly leading to the development of more context-sensitive and semantic-oriented information retrieval models." Note however that we have not attempted here any evaluation of what are the pros and cons, differences and similarities, of our modeling and the other existing approaches, also integrating quantum features. Let us just mention, to give a few examples, Foskett's work in the eighties of last century [25], Agosti et al. work in the nineties [19, 20], and Sordoni et al. more recent work, where the double-slit experiment analogy is also used to investigate quantum interference effects for topic models such as LDA [29].[12]

To conclude, let us observe that in the same way the quantum cognition program, and its effectiveness, does not require the existence of microscopic quantum processes in the human brain [15], the path "towards a quantum Web" that we have sketched here, and in [4], where the Web of written documents is viewed as a "collection of traces" left by an abstract meaning entity—the QWeb—should not be confused with the path "towards a quantum Internet" [23], which is about constructing an Internet able to transmit "quantum information," instead of just "classical information," that is, information carried by entities allowing quantum superposition to also take place and be fully exploited. In the future, there will

[12]Sordoni et al. represented documents as superposition of topics, whereas in our approach documents are considered to be outcomes of the 'tell a story measurements'. In other words, for Sordoni et al. a document is like an electron entering the double-slit apparatus, and the terms like the traces of impact on the detection screen. This is different from our perspective, where documents are instead the traces of impact on the detection screen and the equivalent of the electron entity is the QWeb entity.

certainly be a Quantum Internet and a Quantum Web, that is, there will be a physical Internet more and more similar in structure to the abstract Web of meanings it conveys. These will be fascinating times for the evolution of the human race on this planet, who will then be immersed in a fully developed *noosphere*, but at the moment we are not there yet.

Appendix 1: Interference Plus Context Effects

In this appendix, we show that using the "interference plus context effects" formula (24), all data can in principle be modeled, by suitably choosing the different parameters. For simplicity, we start by assuming that $M^w_X N = N M^w_X$, i.e., that N and M^w_X are compatible, so that the projection $N^\dagger M^w_X N$ can be simply written as $N M^w_X$, as is clear that $N^\dagger M^w_X N = N^\dagger N M^w_X = N^2 M^w_X = N M^w_X$. In other words, we have $(NM^w)^\dagger = NM^w$ and $(NM^w)^2 = NM^w$. This means that we can define the following three orthogonal projectors:

$$P_1 = M^w_X N, \qquad P_2 = (\mathbb{I} - M^w_X)N, \qquad P_3 = \mathbb{I} - N, \tag{26}$$

which are orthogonal to each other:

$$\begin{aligned} P_1 P_2 &= M^w_X N(\mathbb{I} - M^w_X)N = M^w_X N^2 - (M^w_X N)^2 = 0, \\ P_1 P_3 &= M^w_X N(\mathbb{I} - N) = M^w_X N - M^w_X N^2 = 0, \\ P_2 P_3 &= (\mathbb{I} - M^w_X)N(\mathbb{I} - N) = (\mathbb{I} - M^w_X)(N - N^2) = 0. \end{aligned} \tag{27}$$

Consequently, we can write the Hilbert space as the direct sum: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_3$, where $\mathcal{H}_1 = P_1\mathcal{H}$, $\mathcal{H}_2 = P_2\mathcal{H}$, and $\mathcal{H}_3 = P_3\mathcal{H}$ are three orthogonal subspaces, and we can write $|\psi_A\rangle$ and $|\psi_B\rangle$ as linear combinations of vectors belonging to them:

$$\begin{aligned} |\psi_A\rangle &= a e^{i\alpha}|e\rangle + a' e^{i\alpha'}|e'\rangle + a'' e^{i\alpha''}|e''\rangle, \\ |\psi_B\rangle &= b e^{i\beta}|f\rangle + b' e^{i\beta'}|f'\rangle + b'' e^{i\beta''}|f''\rangle, \end{aligned} \tag{28}$$

where $|e\rangle$, $|f\rangle$ are unit vectors in $\mathcal{H}_1$, $|e'\rangle$, $|f'\rangle$ are unit vectors in $\mathcal{H}_2$, and $|e''\rangle$, $|f''\rangle$ are unit vectors in $\mathcal{H}_3$. Considering that the vectors in the expansions (28) are mutually orthogonal, it follows that:

$$\begin{aligned} p_A \mu_A &= \langle\psi_A|N M^w_X N|\psi_A\rangle = \langle\psi_A|P_1|\psi_A\rangle = \langle\psi_A|(a e^{i\alpha}|e\rangle) = a^2, \\ p_B \mu_B &= \langle\psi_B|N M^w_X N|\psi_B\rangle = \langle\psi_B|P_1|\psi_B\rangle = \langle\psi_B|(b e^{i\beta}|f\rangle) = b^2. \end{aligned} \tag{29}$$

We also have:

$$\Re\langle\psi_A|NM^wN|\psi_B\rangle = \Re\langle\psi_A|P_1|\psi_B\rangle = \Re(\langle\psi_A|P_1)(P_1|\psi_B\rangle) = \Re(\langle e|ae^{-i\alpha})(be^{i\beta}|f\rangle)$$
$$= ab\,\Re\, e^{i(\beta-\alpha)}\langle e|f\rangle = abc\,\Re\, e^{i(\gamma+\beta-\alpha)} = abc\cos\phi, \quad (30)$$

where for the second equality we have used $P_1 = P_1^2$, and for the fifth equality we have defined the positive number c and the phase γ such that $c\,e^{i\gamma} = \langle e|f\rangle$, whereas for the last equality we have defined $\phi = \gamma + \beta - \alpha$. In a similar way, we set $c'e^{i\gamma'} = \langle e'|f'\rangle$ and $\phi' = \gamma' + \beta' - \alpha$, and considering that $N = \mathbb{I}N = [M_X^w + (\mathbb{I} - M_X^w)]N = P_1 + P_2$, we have:

$$\Re\,\langle\psi_A|N|\psi_B\rangle = \Re\,\langle\psi_A|P_1|\psi_B\rangle + \Re\,\langle\psi_A|P_2|\psi_B\rangle = abc\cos\phi + a'b'c'\cos\phi'. \quad (31)$$

In a similar way, we have:

$$p_A = \langle\psi_A|N|\psi_A\rangle = \langle\psi_A|P_1|\psi_A\rangle + \langle\psi_A|P_2|\psi_A\rangle = a^2 + a'^2$$
$$p_B = \langle\psi_B|N|\psi_B\rangle = \langle\psi_B|P_1|\psi_B\rangle + \langle\psi_B|P_2|\psi_B\rangle = b^2 + b'^2, \quad (32)$$

from which it follows that:

$$a'^2 = p_A - a^2 = p_A(1-\mu_A) = p_A\bar{\mu}_A, \qquad b'^2 = p_B - b^2 = p_B(1-\mu_B) = p_B\bar{\mu}_B, \quad (33)$$

where we have defined $\bar{\mu}_A = 1 - \mu_A$ and $\bar{\mu}_B = 1 - \mu_B$. We can thus rewrite (24) as:

$$\mu_{AB} = \frac{p_A\,\mu_A + p_B\,\mu_B + 2\sqrt{p_A p_B}\sqrt{\mu_A\mu_B}\,c\cos\phi}{p_A + p_B + 2\sqrt{p_A p_B}\,(\sqrt{\mu_A\mu_B}\,c\cos\phi + \sqrt{\bar{\mu}_A\bar{\mu}_B}\,c'\cos\phi')}. \quad (34)$$

To relate (34) to the webpages' counts, we consider the situation where states are uniform superpositions of states associated with manifest stories (characteristic function states). Different from the "only interference effects situation" of Sect. 4, we however now assume that the vectors represented by characteristic functions are those that are obtained following the action of the context N. Clearly, this should only be considered as a rough approximation meant to illustrate that the present approach can handle the probabilities calculated by performing webpages' counts. So, we assume that $|\psi'_A\rangle = |\chi_A\rangle$ and $|\psi'_B\rangle = |\chi_B\rangle$, so that according to (13), (34) can be written as:

$$\mu_{AB} = \frac{p_A\,\frac{n_{A,X}}{n_A} + p_B\,\frac{n_{B,X}}{n_B} + 2\sqrt{p_A p_B}\sqrt{\frac{n_{A,X}n_{B,X}}{n_A n_B}}\,c\cos\phi}{p_A + p_B + 2\sqrt{p_A p_B}\left(\sqrt{\frac{n_{A,X}n_{B,X}}{n_A n_B}}\,c\cos\phi + \sqrt{\frac{n_{A,X'}n_{B,X'}}{n_A n_B}}\,c'\cos\phi'\right)}, \quad (35)$$

where we have defined $n_{A,X'} = n_A - n_{A,X}$ and $n_{B,X'} = n_B - n_{B,X}$, which are the number of webpages containing the term "A" but not the term "X" and the term "B" but not the term "X," respectively. The consistency of the model is therefore about finding values for $p_A, p_B, c, c' \in [0, 1]$ and $\phi, \phi' \in [0, 2\pi]$, such that (35) can be equal to $\frac{n_{AB,X}}{n_{AB}}$. This will always be the case since (34) can in fact deliver all values between 0 and 1, as we are now going to show.

Consider first the limit case where (34) is equal to 0. Then its numerator has to vanish. If, say, we choose $c = 1$ and $\phi = \pi$, this means that we must have $(\sqrt{p_A\,\mu_A} - \sqrt{p_B\,\mu_B})^2 = 0$, which is satisfied if $\frac{p_A}{p_B} = \frac{\mu_B}{\mu_A}$. For the other limit case where (34) is equal to 1, if we choose $c' = 1$ and $\phi' = \pi$, we have the condition: $(\sqrt{p_A\,\bar{\mu}_A} - \sqrt{p_B\,\bar{\mu}_B})^2 = 0$, which is clearly satisfied if $\frac{p_A}{p_B} = \frac{\bar{\mu}_B}{\bar{\mu}_A}$. For the intermediate values between 0 and 1, if we set $\phi = \phi' = \frac{\pi}{2}$ (no-interference condition), (34) becomes:

$$\mu_{AB} = \frac{p_A}{p_A + p_B}\,\mu_A + \frac{p_B}{p_A + p_B}\,\mu_B, \tag{36}$$

which is a convex combination of μ_A and μ_B. Therefore, by varying p_A and p_B, by just considering context effects all values contained in the interval $[\min(\mu_A, \mu_B), \max(\mu_A, \mu_B)]$ can be obtained.

To be able to extend further the interval, the relative phases ϕ or ϕ' have to be allowed to take values different from $\frac{\pi}{2}$. In this way, also the intervals $[0, \min(\mu(A), \mu(B))]$ and $[\max(\mu(A), \mu(B)), 1]$ can be reached. To see this, we have to study the behavior of $\mu_{AB} = \mu_{AB}(x, x')$ as a function of the two variables $(x, x') = (\cos\phi, \cos\phi')$. We know that $\mu(AB; 0, 0)$ is given by (36), so we just have to show that, for suitable choices of p_A and p_B, by varying x and x' we can reach the 0 value. For a given x, $\mu_{AB}(x, x')$ monotonically decreases as x' increases. Thus, we only have to consider $\mu_{AB}(x, 1)$, and by studying the sign of $\partial_x \mu_{AB}(x, 1)$ one can easily check that (we leave this as an exercise) $\mu_{AB}(x, 1)$ monotonically increases with x. Thus, the minimum corresponds to $\mu_{AB}(-1, 1)$, which is 0 if $c = 1$ and $\frac{p_A}{p_B} = \frac{\mu_B}{\mu_A}$. Similarly, we can consider $\mu_{AB}(x, -1)$ and check that $\mu_{AB}(x, -1)$ also monotonically increases with x. Thus, its maximum corresponds to $\mu_{AB}(1, -1)$, which is 1 if $c' = 1$ and $\frac{p_A}{p_B} = \frac{\bar{\mu}_B}{\bar{\mu}_A}$. In other words, for arbitrary μ_A, μ_B, and μ_{AB}, a quantum representation that can faithfully model the experimental data exist, if both interference and context effects are considered.

Appendix 2: Meaning Bond

In this appendix, we offer a more specific interpretation for the normalized weights a_j characterizing the linear combination in (9), in terms of a notion of *meaning bond* of a concept with respect to another concept, when the QWeb is in a given state $|\psi\rangle$. For this, let M_A and M_B be the projection operators onto the set of QWeb states that are "states of A" and "states of B," respectively. We can then define the ψ-meaning bond $M_\psi(B|A)$ of B towards A by the ratio:

$$M_\psi(B|A) = \frac{p_\psi(B|A)}{p_\psi(B)}, \tag{37}$$

where $p_\psi(B) = \langle\psi|M_B|\psi\rangle$ is the probability for the QWeb's state $|\psi\rangle$ to be successfully tested as being also a "state of B," and

$$p_\psi(B|A) = \frac{\langle\psi|M_A M_B M_A|\psi\rangle}{\langle\psi|M_A|\psi\rangle} \tag{38}$$

is the conditional probability of having the QWeb's state being successfully tested as being a "state of B," when it has been successfully tested to be a "state of A." Indeed, if the QWeb state $|\psi\rangle$ was successfully tested to be a "state of A," according to the *projection postulate* the state immediately following the test is $|\psi_A\rangle = \frac{M_A|\psi\rangle}{\|M_A|\psi\rangle\|}$, which is now a "state of A." And we have $p_\psi(B|A) = \langle\psi_A|M_B|\psi_A\rangle$, hence (38) possesses a sound interpretation as a conditional probability.

The ψ-meaning bond $M_\psi(A|B)$ of A towards B can be similarly obtained by interchanging in (37) the roles of A and B, and since in general $[M_A, M_B] \neq 0$, $M_\psi(A|B) \neq M_\psi(B|A)$, which means that the meaning bond of A towards B will not in general coincide with the meaning bond of B towards A. So, if $p_{\psi_A}(B)$ and $p_\psi(B)$ are interpreted as measuring how much of the meaning of B is present in the QWeb, when the latter is in state $|\psi_A\rangle$ and $|\psi\rangle$, respectively, it is clear that the meaning bond $M_\psi(B|A) = \frac{p_{\psi_A}(B)}{p_\psi(B)}$, being their ratio, it measures the relative increase or decrease of the meaning presence of B when the QWeb state $|\psi\rangle$ is further contextualized by a concept A. In that respect, we can also say that if B is more (resp., less) meaning present in the QWeb, when its state is further contextualized by a concept A, then for such state there is an attractive (resp., repulsive) meaning bond of B towards A, whereas if $p_{\psi_A}(B) = p_\psi(B)$ the meaning bond can be said to be neutral. Also, since we have $p_{\psi_B}(B) = 1$, the meaning bond of B towards itself is $M_\psi(B|B) = p_\psi^{-1}(B)$, so that there will be self-neutrality when $p_\psi(B) = 1$, and self-attraction if $p_\psi(B) < 1$ (but there cannot be self-repulsion).

We now observe that: $p_\psi(W_j)M_\psi(W_j|A) = p_\psi(W_j|A) = \langle\psi_A|P_j|\psi_A\rangle = a_j^2$, where $P_j = |e_j\rangle\langle e_j|$ is the projection operator onto the one-dimensional subspace generated by the 'ground state of W_j," i.e., of the story-concept indicated by the specific combination of words contained in the webpage W_j. Thus, we have that the coefficients a_j in the expansion of the state $|\psi_A\rangle = \frac{M_A|\psi\rangle}{\|M_A|\psi\rangle\|} = \sum_{j=1}^n a_j e^{i\alpha_j}|e_j\rangle$, which is a "state of A," can be written as:

$$a_j = \sqrt{p_\psi(W_j)M_\psi(W_j|A)} \tag{39}$$

and therefore are given by (the square root of) the "ψ-meaning bond of W_j towards A," normalized by the probability $p_\psi(W_j)$, and in that sense we can say that they express a meaning connection between A and the W_j. Note also that in the case where $|\psi\rangle$ corresponds to the uniform state $|\chi\rangle = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} e^{i\rho_j} |e_j\rangle$, (37) reduces to the ratio

$$M_\chi(B|A) = \frac{n\, n_{AB}}{n_A n_B}, \tag{40}$$

which corresponds to the more specific notion of meaning bond introduced in [3] (see also [5]).

Acknowledgement This work was supported by the Marie Skłodowska-Curie Innovative Training Network 721321—"QUARTZ."

References

1. Aerts, D. (2005). Towards a new democracy: Consensus through quantum parliament. In D. Aerts, B. D'Hooghe & N. Note (Eds.), *Worldviews, science and us, redemarcating knowledge and its social and ethical implications* (pp. 189–202). Singapore: World Scientific.
2. Aerts, D. (2009). Quantum structure in cognition. *Journal of Mathematical Psychology, 53*, 314–348.
3. Aerts, D. (2011). Measuring meaning on the world-wide web. In D. Aerts, J. Broekaert, B. D'Hooghe & N. Note (Eds.), *Worldviews, science and us: Bridging knowledge and its implications for our perspectives of the world* (pp. 304–313). Singapore: World Scientific.
4. Aerts, D., Argüelles, J., Beltran, L., Beltran, L., Geriente, S., Sassoli de Bianchi, M., et al. (2018). Towards a quantum world wide web. *Theoretical Computer Science*. https://doi.org/10.1016/j.tcs.2018.03.019
5. Aerts, D., Argüelles, J., Beltran, L., Beltran, L., Sassoli de Bianchi, M., Sozzo, S., et al. (2017). Testing quantum models of conjunction fallacy on the world wide web. *International Journal of Theoretical Physics, 56*, 3744–3756.
6. Aerts, D., Argüelles, J., Beltran, L., Geriente, S., Sassoli de Bianchi, M., Sozzo, S., et al. (2018). Spin and wind directions I: Identifying entanglement in nature and cognition. *Foundations of Science, 23*, 323–335. https://doi.org/10.1007/s10699-017-9528-9
7. Aerts, D., Argüelles, J., Beltran, L., Geriente, S., Sassoli de Bianchi, M., Sozzo, S., et al. (2018). Spin and wind directions II: A bell state quantum model. *Foundations of Science, 23*, 337–365. https://doi.org/10.1007/s10699-017-9530-2
8. Aerts, D., & Sassoli de Bianchi, M. (2014). The extended Bloch representation of quantum mechanics and the hidden-measurement solution to the measurement problem. *Annals of Physics, 351*, 975–102. https://doi.org/10.1016/j.aop.2014.09.020. See also the erratum: *Annals of Physics, 366*, 197–198. https://doi.org/10.1016/j.aop.2016.01.001
9. Aerts, D., & Sassoli de Bianchi, M. (2015). The unreasonable success of quantum probability I: Quantum measurements as uniform measurements. *Journal of Mathematical Psychology, 67*, 51–75.
10. Aerts, D., & Sassoli de Bianchi, M. (2016). The extended Bloch representation of quantum mechanics. Explaining superposition, interference and entanglement. *Journal of Mathematical Physics, 57*, 122110. https://doi.org/10.1063/1.4973356
11. Aerts, D., & Sassoli de Bianchi, M. (2017). *Universal measurements*. Singapore: World Scientific.
12. Aerts, D., & Sassoli de Bianchi, M. (2017). Beyond-quantum modeling of question order effects and response replicability in psychological measurements. *Journal of Mathematical Psychology, 79*, 104–120. https://doi.org/10.1016/j.jmp.2017.03.004

13. Aerts, D., Sassoli de Bianchi, M., & Sozzo, S. (2016). On the foundations of the Brussels operational-realistic approach to cognition. *Frontiers in Physics, 4*, 17. https://doi.org/10.3389/fphy.2016.00017
14. Aerts, D., Sassoli de Bianchi, M., Sozzo, S., & Veloz, T. (2018). On the conceptuality interpretation of quantum and relativity theories. *Foundations of Science*. https://doi.org/10.1007/s10699-018-9557-z
15. Aerts, D., Sassoli de Bianchi, M., Sozzo, S., & Veloz. T. (2018). Modeling human decision-making: An overview of the Brussels quantum approach. *Foundations of Science*. https://doi.org/10.1007/s10699-018-9559-x
16. Aerts, D., & Sozzo, S. (2016). Quantum structure in cognition: Origins, developments, successes and expectations. In E. Haven & A. Khrennikov (Eds.), *The Palgrave handbook of quantum models in social science: Applications and grand challenges* (pp. 157–193). London: Palgrave & Macmillan.
17. Aerts, D., Sozzo, S., & Veloz, T. (2015). Quantum structure in cognition and the foundations of human reasoning. *International Journal of Theoretical Physics, 54*, 4557–4569.
18. Aerts, D., Sozzo, S., & Veloz, T. (2015). Quantum structure of negation and conjunction in human thought. *Frontiers in Psychology, 6*, 1447. https://doi.org/10.3389/fpsyg.2015.01447
19. Agosti, M., Colotti, R., & Gradenigo, G. (1991). A two-level hypertext retrieval model for legal data. In *SIGIR '91: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 316–325). New York, NY: ACM Press.
20. Agosti, M., Crestani, F., & Melucci, M. (1995). Automatic authoring and construction of hypermedia for information retrieval. *ACM Multimedia Systems, 3*, 15–24.
21. Bach, R., Pope, D., Liou, S.-H., & Batelaan, H. (2013). Controlled double-slit electron diffraction. *New Journal of Physics, 15*, 033018.
22. D'Ariano, G. M., Paris, M. G. A., & Sacchi, M. F. (2004). Quantum tomographic methods. *Lecture Notes in Physics, 649*, 7–58.
23. Dür, W., Lamprecht, R., & Heusler, S. (2017). Towards a quantum internet. *European Journal of Physics, 38*, 043001.
24. Feynman, R. P. , Leighton, R. B., & Sands, M. (1964). *The Feynman Lectures on Physics* (Vol. III), New York, NY: Addison-Wesley (revised and extended edition in 2005).
25. Foskett, D. (1980). Thesaurus. In A. Kent, H. Lancour & J. Daily (Eds), *Encyclopedia of library and information science* (Vol. 30, pp. 416–462). New York, NY: Marcel Dekker.
26. Melucci, M. (2012). Contextual search: A computational framework. *Foundation and Trends in Information Retrieval, 6*, 257–405.
27. Melucci, M. (2015). Introduction to information retrieval and quantum mechanics. In *The Information Retrieval Series* (Vol. 35). Berlin: Springer. https://doi.org/10.1007/978-3-662-48313-8
28. Sassoli de Bianchi, M. (2015). God may not play dice, but human observers surely do. *Foundations of Science, 20*, 77–105. https://doi.org/10.1007/s10699-014-9352-4
29. Sordoni, A., He, J., & Nie, J.-Y. (2013). Modeling latent topic interactions using quantum interference for information retrieval. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management - CIKM '13* (pp. 1197–1200). New York, NY: ACM Press. https://doi.org/10.1145/2505515.2507854
30. Van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge: Cambridge University Press.

Non-separability Effects in Cognitive Semantic Retrieving

Aleksey V. Platonov, Igor A. Bessmertny, Evgeny K. Semenenko, and Alexander P. Alodjants

Abstract This paper discusses a Bell test analogue known in quantum physics, which allows determining the presence of non-separability features by using semantic search of information and document ranking for articles in Russian. The model of Bell test in semantics is based on hyperspace analogue to language (HAL) algorithm provides to obtain vector representation of words (in Hilbert space) using the dictionary index and considering the word order. We show the existence of certain quantum-like correlations between two words of the user's query; these correlations cannot be taken into account in the classical probabilistic description. We predict that the contextuality revealed can be regarded as human cognitive level both while writing of certain texts and queries to them.

Keywords Information retrieval systems · Decision-making · Quantum cognitive science · Quantum entanglement · Contextuality · Machine learning

1 Introduction

Over the past decade, the rapid growth of information resources in terms of information transmission and processing has led to an exponential growth of data, most of which are poorly structured or have no structure at all. The need to proceed and analyze such data in real time is a serious problem, which is directly related to the safety of society itself in various areas of economics, finances, and social sphere. Thematic modeling as one of the machine learning paradigms is an important tool for modern text and document analysis and has a direct application to the problems of information retrieval. On the other hand, quantum approach for information retrieving allows to take into account some peculiarities (disturbances) occurring during the "interaction" of the user and "smart" search system similarly to

A. V. Platonov · I. A. Bessmertny · E. K. Semenenko · A. P. Alodjants (✉)
ITMO University, Saint Petersburg, Russia

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_2

quantum measurement paradigm [10]. As a result, it should lead to a more accurate formulation of requests context and, hence, to finding higher relevance of issued documents. In this regard, the model of semantic space looks very limited, as it is based on the so-called bag-of-words approach without word order consideration, when the meaning is coded by the counters of words included in the context of the object.

Quantum cognition represents one of the modern approaches taking into account the contextuality of interaction of the user and smart information system [2, 5]. Interestingly, in Ref. [4] there were the first attempts to describe psychological aspects in decision-making via quantum probabilistic methods and quantum measurement theory approaches.

Nowadays, there is an increased interest in the application of quantum formalism to information retrieval problems, see, e.g., [6, 8, 9, 12, 13]. In particular, it is shown that information retrieval models such as logical, probabilistic, and vector ones, can be described with quantum formalism in Hilbert space. As a result, it is possible to take into account the contextuality of queries [8]. Notably, in [13] the authors formulate the quantum probability ranking principle, which is a generalization of a well-known probability ranking principle used to assess the criteria for ranking of issued documents considering links between documents. At the same time, to obtain the best overall search efficiency, the information system ranks documents not only in descending order of the probability of their relevance to the user, but also considering the effects of "quantum interference." As shown in [11, 12], the analogies with quantum measurement of photon polarization considering in the framework of quantum optics paradigm are relevant here, cf. [1].

The aim of this paper is to demonstrate a quantum-like ranking algorithm using a Bell test performed with text samples given in Russian taking into account the contextuality and compatibility of different queries.

2 Bell Test in the Problem of Cognitive Semantic Information Retrieval

2.1 Bell Inequality and Its Interpretation

In the problem under consideration, the most suitable toolkit is quantum theory, which originally aimed to simulate incompatible physical experiments. In such experiments a specific experimental configuration (context) allows to consistently determine a certain set of physical quantities, which become undefined in another experimental configuration (context). In quantum cognitive science, such incompatible experimental situations correspond to incompatible cognitive contexts, the use of which in decision-making leads to violations of classical (Boolean) logic. A consequence of quantum theory is the possible occurrence of correlations between measurement results (conducted in physics over remote systems), which are stronger than allowed by classical theories with hidden parameters [7, 10].

Suppose there are two subsystems A and B, and experiments are performed, respectively, each having two possible outcomes coded by values of ± 1. This condition is satisfied, for example, by the system of electron pairs, whose spins are measured. Each of experiments A and B is performed in two versions A and A $'$, B and B$'$, which differ in the direction of spin measurement: 0°, 90°, 45°, and 135°, respectively. These directions are combined in four possible ways: $\{A, B\}$, $\{A, B'\}$, $\{A', B\}$, $\{A', B'\}$ in each configuration to obtain a statistically significant set of (probabilistic) outcomes. The so-called Clauser–Horne–Shimony–Holt (CHSH) type Bell *inequality*

$$S = \left|\langle AB\rangle - \langle AB'\rangle + \langle A'B\rangle + \langle A'B'\rangle\right| \leq 2 \tag{1}$$

is used to distinguish purely quantum and classical correlations in theory. Inequality (1) holds for separable states, if factorization for statistically average values is true, i.e., $\langle \ldots \rangle = \langle \ldots \rangle \langle \ldots \rangle$.

In quantum theory, Tsirelson's bound predicts the existence of quantum states of two subsystems for which (see [3])

$$2 < S \leq 2\sqrt{2}. \tag{2}$$

In physics, these states are called entangled or non-separable [7]. Violation of CHSH inequality provides evidence for contextual interdependence for a given set of systems and experimental procedures.

2.2 Bell Test in Semantic Retrieving

In this paper, we use a Bell test, the experiment that checks inequalities discussed above, and conducted with objects in the semantic Hilbert space. In our experiment we use hyperspace analogue to language (HAL) algorithm for obtaining the vector representation of words, which unlike the popular bag of words allows to consider the word order in a sentence and thereby increasing the accuracy of determining dependencies between words. The HAL algorithm also gives the vector representation of words based on the vocabulary index (the word matches some unique numeric identifier) and the set of documents processed. To obtain such a representation, the matrix is built with cells containing sums of distances between the word corresponding to a row and the word corresponding to a column in the frame of the text corpus. At the same time, the relationship distance from a word in a row to a word in a column if the word in the column is to the rightİ is modeled. Thus, the word order in the sentence is taken into account. Moreover, the distance is not calculated between all the words in the matrix, but only between those pairs of words that are closer to each other than a predetermined distance, which is called the HAL window size.

To perform the Bell test we use text files obtained from several Wikipedia articles in Russian on Programming Language: Programming Language, Programming, Java, C++. The content of these articles without the layout was subjected to preprocessing. While reading the files, a text index is constructed. It contains the normalized forms of words and their correspondence to a unique numeric identifier. This numeric identifier is used to determine one vector coordinate in the vector space of words obtained by HAL algorithm. After obtaining the index of the words for the document being processed, HAL matrix is built (see the next subsection). This allows to get a vector representation of the words of the document, and the average value of the sum of these vectors is the document vector. Query word vectors are derived from the same HAL matrix. The resulting document and query vectors are used to calculate a Bell parameter.

Thus, as far as the index and vector representations of words are obtained, the words of the user query, such as Programming Language, are reduced to a vector form and then normalized. On these words two bases are constructed defining the subspaces of vectors corresponding to the two query words. To obtain a basis, the Schmidt orthogonalization algorithm is used. The document vector decomposes according to these bases:

$$\left|D_{w_1}\right\rangle = a\left|+\right\rangle_A + b\left|-\right\rangle_A ; \left|D_{w_2}\right\rangle = c\left|+\right\rangle_B + d\left|-\right\rangle_B , \tag{3}$$

where $\left|D_{w_1}\right\rangle$ is the document vector; $|+\rangle_A$ ($|-\rangle_A$) and $|+\rangle_B$ ($|-\rangle_B$) are bases in which queries A and B are fully relevant (not relevant) to the document, respectively. Next, we define (projective) measurement operators for queries A and B, respectively:

$$\begin{aligned} A_x &= |+\rangle_{AA}\langle -| + |-\rangle_{AA}\langle +| ; A_z = |+\rangle_{AA}\langle +| - |-\rangle_{AA}\langle -| ; \\ B_x &= |+\rangle_{BB}\langle -| + |-\rangle_{BB}\langle +| ; B_z = |+\rangle_{BB}\langle +| - |-\rangle_{BB}\langle -| . \end{aligned} \tag{4}$$

The operators related to the same queries (for example, to A or to B) do not commute with each other and correspond to operators of measurement of spin projections in quantum physics [7, 10]. Operator A_z allows us to determine the degree of the document relevance with the first word of the query. In this case, if the document vector in the basis of the first word is represented by the vector $[a, b]$ (3), then the probability that the document relates to the subject of the first word can be obtained by the Born rule$\langle A_z \rangle_D = \langle D | A_z | D \rangle = a^2 - b^2$.

We also use the combinations $B_+ = (B_z - B_x)/\sqrt{2}$, $B_- = -(B_z + B_x)/\sqrt{2}$ which determine the degree of correspondence of the document to the second word in the rotated basis. Note that the calculations can be performed in the basis of one of the words, for example, A. The resulting set of operators is used to calculate the Bell parameter of queries:

$$S_q = |\langle A_z B_+ \rangle + \langle A_x B_+ \rangle| + |\langle A_z B_- \rangle - \langle A_x B_- \rangle| . \tag{5}$$

3 Results

Figure 1 displays the results of the Bell test performed with Russian texts on Language (word A) of Programming (word B) from Wikipedia. As seen, the Bell test essentially depends on the size of the HAL window under the modeling semantic space. For a large HAL window all four curves approach the value of quantum limit $2\sqrt{2}$. This can be explained by the fact that with increasing the size of the window, the terms included in the user's request overlap more often with each other contexts. The same can be said about the values of less than 2. However, it is important that all the curves in Fig. 1 reach the quantum entanglement domain at different window sizes. From Fig. 1 it is evident that the document "Programming Language" first reaches this domain as it directly relates to the subject of the request. Then the documents "Java," "C++," and "Programming" follow.

Thus, we can conclude that a Bell test can be used as a sign of relevance (contextuality) of the article to the subject of a request.

The bigger value of the Bell test parameter does not mean the greater entanglement. However, the size of a HAL window, which "quantum entanglement" or "inseparability" appears on, can indicate the presence of a subject searched in an article.

Moreover, this test can be used to separate text documents by means of request subjects. Documents more relevant to a query subject reveal Bell test parameter value greater than 2 with less size of a HAL window. This interesting feature can be applied in two ways.

First, one can rank the documents by a size of a HAL window giving Bell test parameter greater than 2. More relevant documents have less size of a window.

Second, this test may be used for subject terms extraction. Choosing a fixed window size one can extract pairs of word chains for which a Bell test parameter is greater than a threshold. Further we are going to apply this test for more extensive testing of these hypotheses.

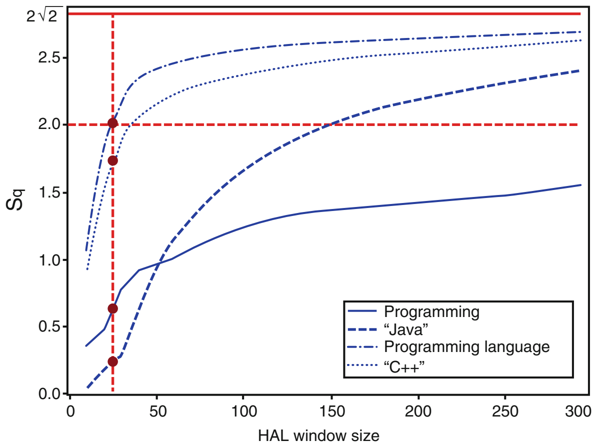


Fig. 1 Bell test parameter vs. HAL window size for four documents. The dots characterize Bell parameter values indicating the ranking of the documents at the request

Acknowledgements This work was financially supported by the Government of the Russian Federation, Grant 08-08. The authors are grateful to Prof. Andrei Khrennikov and Dr. Ilya Surov for fruitful discussions.

References

1. Alodjants, A. P., & Arakelian, S. M. (1999). Quantum phase measurements and nonclassical polarization states of light. *Journal of Modern Optics, 46*(3), 475–507.
2. Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision* (p. 408). Cambridge: Cambridge University Press.
3. Cirel'son, B. S. (1980). Quantum Generalizations of Bell's Inequality. *Letters in Mathematical Physics, 4*, 93.
4. Gurevich, I. I., & Feigenberg, I. M. (1977). *What probabilities work in psychology? Probabilistic forecasting of human activity* (pp. 9–21). Moscow: Science.
5. Khrennikov, A. Yu. (2010). *Ubiquitous quantum structure. From psychology to finance* (p. 216). Berlin: Springer.
6. Melucci, M. (2015). *Introduction to information retrieval and quantum mechanics. The information retrieval series* (Vol. 35, p. 232). Berlin: Springer.
7. Peres, A. (2002). *Quantum theory: Concepts and methods* (p. 464). Dordrecht: Kluwer.
8. Piwowarski, B., Frommholz, I., Lalmas, M., & van Rijsbergen, C. J. (2010). What can quantum theory bring to information retrieval. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 59–68). New York, NY: ACM.
9. Van Rijsbergen, C. J. (2004). *The geometry of information retrieval* (p. 150). New York, NY: Cambridge University Press.
10. Wheeler, J. A., & Zurek W. (Eds.) (1983). *Quantum theory and measurement* (p. 811). Princeton, NJ: Princeton University Press.
11. Zhang, P., Li, J., Wang, B., Zhao, X., Song, D., Hou, Yu., et al. (2016). A quantum query expansion approach for session search. *Entropy, 18*, 146.
12. Zhang, P., Song, D., Zhao, X., & Hou, Y. (2011). Investigating query-drift problem from a novel perspective of photon polarization. in *Proceedings of the Conference on the Theory of Information Retrieval* (pp. 332–336). Berlin: Springer.
13. Zuccon, G., & Azzopardi, L. (2010). Using the quantum probability ranking principle to rank interdependent documents. In: *Proceedings of the European Conference on Information Retrieval* (pp. 357–369). Berlin: Springer.

Introduction to Hilbert Space Multi-Dimensional Modeling

Jerome Busemeyer and Zheng Joyce Wang

Abstract This chapter provides a brief introduction to procedures for estimating Hilbert space multi-dimensional (HSM) models from data. These models, which are built from quantum probability theory, are used to provide a simple and coherent account of a collection of contingency tables. The collection of tables are obtained by measurement of different overlapping subsets of variables. HSM models provide a representation of the collection of the tables in a low dimensional vector space, even when no single joint probability distribution across the observed variables can reproduce the tables. The parameter estimates from HSM models provide simple and informative interpretation of the initial tendencies and the inter-relations among the variables.

Keywords Quantum probability · Non-commutativity · Data fusion

1 Introduction

This chapter provides an introduction to computational tools, based on what we call Hilbert space multi-dimensional theory, which can be used for representing data tables from multiple sources by a single coherent vector space and linear operations on the space. For more complete descriptions of this theory, see the original articles by the authors Busemeyer and Wang [8, 9], which include detailed worked out examples. Here we plan to outline the main steps of building a program and also point to computer programs available to process collections of data tables.

J. Busemeyer (✉)
Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
e-mail: jbusemey@indiana.edu

Z. J. Wang
School of Communication, Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH, USA

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_3

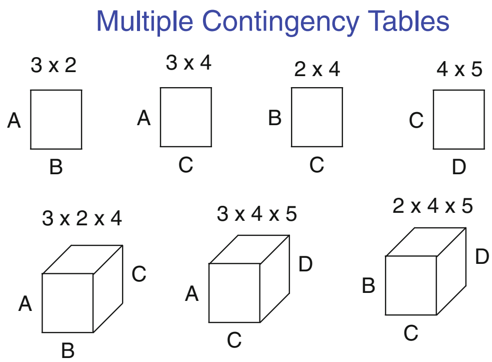


Fig. 1 Illustration of a collection of contingency data tables

Figure 1 illustrates the basic problem that we wish to address. Suppose large medical data sites provide information about co-occurrence of various kinds of symptoms, labeled A, B, C, and D in the figure. The symptoms can be manifest to different degrees. For example, symptom B is binary valued, symptom A has three levels, symptom C has four degrees, and symptom D has five rating values. Suppose different methods for querying the sites yield different contingency tables summarizing co-occurrence of pairs of variables and co-occurrence of triples of variables, which produce the tables shown in the figure. The cells of the contingency tables contain the frequency of a combination of symptoms. For example, the A by B by C table is a 3 by 2 by 4 table, and each cell contains the frequency that a particular combination of values was assigned to the variables A, B, C, using one query method.

The following problem arises from considering all these contingency tables. How does a data scientist integrate and synthesize these seven different tables into a compressed, coherent, and interpretable representation? This is a problem that often arises in relational database theory [1]. It is common to apply categorical data analysis [3] to a single table (e.g., a single A by B by C by D table). However, the problem is different here because there are a collection of seven tables of varying dimensions rather than a single four-way table. Alternatively, one could try Bayesian networks, which require assuming that all the tables are generated from a *single* latent four-way joint distribution [10]. Unfortunately, however, it may be the case that no four-way joint distribution exists that can reproduce all the observed tables! This occurs when the data tables violate consistency constraints, forced by marginalization, upon which Bayes nets rely to fit the tables [11].

Hilbert space multi-dimensional (HSM) models are based on quantum probability theory [13]. They provide a way to account for a collection of tables, such as illustrated in Fig. 1, even when no four-way joint distribution exists. HSM models provide an estimate of the target population's initial tendencies in the form of a state vector, and HSM models represent the inter-relationships between the different variables (symptoms in this example) using "rotations" of the basis of the vector space.

This chapter is organized as follows: First, we summarize the basic principles of quantum probability theory, then we summarize the steps required to build an HSM model, and finally we refer to programs available on the web for applying an HSM model to real data.

2 Basics of Quantum Probability Theory

The idea of applying quantum probability to the field of judgment began from several directions [2, 4, 5, 14]. The first to apply these ideas to the field of information retrieval was by van Rijsbergen [18]. For more recent developments concerning the application of quantum theory to information retrieval, see [16]. van Rijsbergen argues that quantum theory provides a sufficiently general yet rigorous formulation for integration of three different approaches to information retrieval—logical, vector space, and probabilistic. Another important reason for considering quantum theory is that human judgments (e.g., judging presence of symptoms by doctors) have been found to violate rules of Kolmogorov probability, and quantum probability provides a formulation for explaining these puzzling findings (see, e.g., [7]).

HSM models are based on quantum probability theory and so we need to briefly review some of the basic principles used from this theory.[1]

In quantum theory, a variable (e.g., variable A) is called an observable, which corresponds to the Kolmogorov concept of a random variable. The pair of a measurement of a variable and an outcome generated by measuring a variable is an event (e.g., measurement of variable A produces the outcome 3, so that we observe $A = 3$).

Kolmogorov theory represents events as subsets of a sample space, Ω. Quantum theory represents events as subspaces of a Hilbert space H.[2] Each subspace, such as A, corresponds to an orthogonal projector, denoted P_A for subspace A. An orthogonal projector is used to project vectors into the subspace it represents.

In Kolmogorov theory, the conjunction "A and B" of two events, A and B, is represented by the intersection of the two subsets representing the events (e.g., $(A = 3) \cap (B = 1)$). In quantum theory, a sequence of events, such as A and then B, denoted AB, is represented by the sequence of projectors $P_B P_A$. If the projectors commute, $P_A P_B = P_B P_A$, then the product of the two projectors is a projector corresponding to the subspace $A \cap B$, that is, $P_B P_A = P(A \cap B)$; and the events A and B are said to be *compatible*. However, if the two projectors do not commute, $P_B P_A \neq P_A P_B$, then neither their product is a projector, and the events

[1] See [7, 15, 16, 18] for tutorials for data and information scientists.

[2] Technically, a Hilbert space is a complex valued inner product vector space that is complete. Our vectors spaces are finite, and so they are always complete.

are *incompatible*. The concept of incompatibility is a new contribution of quantum theory, which is not present in Kolmogorov theory.

Kolmogorov theory defines a state as a probability measure p that maps events to probabilities. Quantum theory uses a unit length state vector, here denoted ψ, to assign probabilities to events.[3] Probabilities are then computed from the quantum algorithm

$$p(A) = \| P_A \psi \|^2 . \tag{1}$$

Both Kolmogorov and quantum probabilities satisfy the properties for an additive measure. In the Kolmogorov case, $p(A) \geq 0$, $p(\Omega) = 1$, and if $(A \cap B) = 0$, then $p(A \cup B) = p(A) + p(B)$.[4] In the quantum case, $p(A) \geq 0$, $p(H) = 1$, and if $P_A P_B = 0$, then $p(A \vee B) = p(A) + p(B)$.[5] In fact, Eq. (1) is the unique way to assign probabilities to subspaces that form an additive measure for dimensions greater than 2 [12].

Kolmogorov theory defines a conditional probability function as follows:

$$p(B|A) = \frac{p(A \cap B)}{p(A)},$$

so that the joint probability equals $p(A \cap B) = p(A)p(B|A) = p(B)p(A|B) = p(B \cap A)$, and order does not matter. In quantum theory, the corresponding definition of a conditional probability is

$$p(B|A) = \frac{\| P_B P_A \psi \|^2}{p(A)},$$

and so the probability of the sequence AB equals $p(AB) = p(A) \cdot p(B|A) = \| P_B P_A \psi \|^2$. In the quantum case, this definition of conditional probability incorporates the property of incompatibility: if the projectors do not commute, so that $P_A P_B \neq P_B P_A$, then $p(AB) \neq p(BA)$, and order of measurement matters. Extensions to sequences with more than two events follow the same principles for both classical and quantum theories. For example, the quantum probability of the sequence $(AB)C$ equals $\| P_C \, (P_B P_A) \, \psi \|^2$.

[3] A more general approach uses what is called a density operator rather than a pure state vector, but to keep ideas simple, we use the latter.

[4] $\cup$ is the union of subsets A, B.

[5] $\vee$ is the span of subspaces A, B.

3 Steps to Build an HSM Model

An HSM model is constructed from the following six steps:

1. Determine the compatibility and incompatibility relations among the variables.
2. Determine the dimension of the Hilbert space based on assumed compatibility relations.
3. Define the initial state given the dimension of the Hilbert space.
4. Define the projectors for the variables using unitary transformations to change the basis.
5. Compute the choice probabilities given the initial state and the projectors.
6. Estimate model parameters, compare fit of models.

3.1 How to Determine the Compatibility Relations

There are two ways to investigate and determine compatibility between a pair of variables. The direct way is to empirically determine whether or not the joint frequencies change depending on order of presentation. If there are order effects, then that is evidence for incompatibility. An indirect way is to compare fits of models that assume different compatibility relations. This indirect methods might be needed if no empirical tests of order effects are available.

3.2 How to Determine the Dimension

The basic idea of HSM modeling is to start with the minimum dimension required, and then add dimensions only if needed to obtain a satisfactory fit to the data. Of course this model comparison and model selection process needs to provide a reasonable balance between accuracy and parsimony. For example, when fitting the models using maximum likelihood estimation, model comparison indices such as Bayesian information criterion or Akaike information criterion can be used.

The minimum dimension is determined by the maximum number of combinations of values produced by the maximum number of compatible variables. For example, in Fig 1, suppose variables B and C are compatible with each other, and variables A and D are compatible with each other, but the pair B,C is incompatible with the pair A,D. In this case, there are at most two compatible variables. The B,C pair produces $2 \cdot 4 = 8$ combinations of values, but the A,D pair produces $3 \cdot 5 = 15$ combinations. The minimum dimension needs to include all 15 combinations produced by the A,D pair. Therefore, the minimum dimension is 15 in this example.

3.3 Define the Initial State

The compatible variables can be chosen to form the basis used to define the coordinates of the initial state ψ. In this example, the space is 15 dimensional, and the compatible pair, A,D can be chosen to define the initial basis for the unit length 15×1 column matrix ψ. Each coordinate ψ_{ij} represents the amplitude corresponding to the pair of values $(A = i, D = j)$, $i = 1, 2, 3$; $j = 1, 2, \ldots, 5$, for representing the initial state. The squared magnitude of a coordinate equals the probability of the combination, $p(A = i, D = j) = \left|\psi_{ij}\right|^2$. In general, the coordinates can be complex, but in practice they are usually estimated as real values.

3.4 Define the Projectors

The orthogonal projector for an event that is defined in the initial basis is simply an indicator matrix that picks out the coordinates that correspond to the event. For example, using the previous example, the projector for the event $(A = i, D = j)$ is simply $P_{A=i,D=j} = diag\left[0 \cdots 1 \cdots 0\right]$, where the one is located in the row corresponding to (i, j), which is a one-dimensional projector. The projector for the event $(A = i)$ equals $P_{A=i} = \sum_j P_{A=i,D=j}$ and the projector for the event $(D = j)$ equals $P_{D=j} = \sum_i P_{A=i,D=j}$, and these are multi-dimensional projectors.

The projectors for the incompatible events require changing the basis from the original basis to the new basis for the incompatible variables. For example, suppose we wish to define the events for the variables B,C. If we originally defined the initial state ψ in the B,C basis from the start, then these projectors would simply be defined by indicator matrices as well. Recall that the dimension of the space is 15, and there are only 8 combination of values for B, C. Therefore one or more of the combinations for B, C need to be defined by a multi-dimensional projector, M_{kl}, which is simply an indicator matrix, such as $M_{kl} = diag\left[1\ 0 \ldots 1\ 0\right]$ that picks out two or more coordinates for the event $(B = k, C = l)$. The collection of indicator matrices, $\{M_{kl}, k = 1, 2; l = 1, 2, 3, 4\}$, forms a complete orthonormal set of projectors. The projector for the event $(B = k)$, in the B, C basis, is simply $p(B = k) = \sum_l M_{B=k,C=l}$, and the projector for $(C = l)$ in the B, C basis is $p(C = l) = \sum_k M_{B=k,C=l}$.

We did not, however, define the initial state ψ in terms of the B,C basis. Instead we defined the initial state ψ in terms of the A,D basis. Therefore we need to "rotate" the basis from the A, D basis to the B, C basis to form the B,C projectors as follows: $P_{B=k,C=l} = U \cdot M_{jk} \cdot U^\dagger$, where U is a unitary matrix (an orthonormal matrix). Now the projector for the event $(B = k)$, in the A, D basis, is $p(B = k) = \sum_l P_{B=k,C=l}$, and the projector for $(C = l)$ in the A, D basis is $p(C = l) = \sum_k P_{B=k,C=l}$.

Any unitary matrix can be constructed from a Hermitian matrix $H = H^\dagger$ by the matrix exponential $U = exp(-i \cdot H)$. Therefore, the most challenging problem is to construct a Hermitian matrix that captures the change in bases. This is facilitated by using substantive theory from the domain under investigation. We describe this step in more detail in the original articles.

3.5 Compute the Choice Probabilities

Once the projectors have been defined, it is straightforward to compute the probabilities for any contingency table using the quantum algorithm described earlier. For example, the probabilities for the AB table are obtained from the equation $p(A = i, B = j) = \left\| P_{B=j} P_{A=i} \psi \right\|^2$, and the probabilities for the A, B, D table are obtained from the equation $p(A = i, B = k, D = j) = \left\| P_{D=j} P_{B=k} P_{A=i} \cdot \psi \right\|^2$.

3.6 Estimate Model Parameters, Compare and Test Models

Once the model has been defined, the parameters of the initial state ψ and the parameters in the Hamiltonian matrix H can be estimated from the frequencies contained within contingency table data. This can be accomplished by using maximum likelihood estimation procedures. Suppose the dimension equals n ($n = 15$ in our example). If we use a real valued initial state, then initial state has $n - 1$ parameters (because the state is restricted to unit length). If the Hamiltonian is restricted to real values, then the Hamiltonian has $(n \cdot (n + 1)/2) - 1$ parameters (one diagonal entry is arbitrary). However, often it is possible to use a lower number of parameters for the Hamiltonian. Model comparison methods, such as Bayesian information criterion or Akaike information criterion, can be used to compare models for accuracy adjusted for parsimony (defined by number of parameters).

HSM models can also be empirically tested using a generalization criterion. After estimating the projectors from two-way tables shown in Fig. 1, the model can be used to make a priori predictions for table A by D or for a three-way table such as A by B by D. This provides strong empirical tests of the model predictions.

The model also provides interpretable parameters to help understand the complex array of contingency tables. The estimate of the initials state ψ provides the initial tendencies to respond to questions. In the previous example, ψ represents the probabilities to respond to the A, D questions. The rotation, $U^{\dagger} \psi$ gives the initial tendencies to respond to the B, C questions. The squared magnitude of an entry in the unitary matrix, $\left| u_{jk} \right|^2$, represents the squared correlation between a basis vector representing an event in one basis (e.g., an event in the A, D basis) and a basis vector representing an event in another basis (e.g., the B, C basis). These squared correlations describe the inter-relations between the variables, independent of the initial state.

4 Computer Programs

We have started developing computer programs for fitting HSM models to different collections of contingency tables. These programs are currently located at the following site http://mypage.iu.edu/~jbusemey/quantum/HilbertSpaceModelPrograms.htm.

The site contains a link to some commonly used programs required for all of the models. It also contains programs designed to fit (a) collections of one and two-way tables made from binary variables, such as those that appear in [9], (b) one and two-way tables for variables with 2, 3, 4 values, such as those that appear in [8], (c) a model for order effects between a pair of variables with a relatively large (e.g., nine or greater) levels of rating scale values, such as those that appear in [19].

5 Concluding Comments

HSM models provide a simple and low dimensional method for representing multiple contingency tables formed from measurement of subsets of variables. This simple representation in low dimensional spaces is achieved by using “rotation” of the basis vectors to generate new incompatible variables. Bayesian network models can also be applied to collections of tables; however, these types of models assume the existence of a complete joint distribution of the observed variables, and it is often the case that no complete joint distribution can reproduce the tables because of violations of constraints imposed by marginalization. HSM models can be applied to collections of tables even when no complete joint distribution exists to reproduce the collection. Of course, HSM models do not provide the only way, and there are other probabilistic models that could be considered such as the use of probabilistic data base programming methods [6]. However, HSM models have been shown to provide successful accounts of actual empirical data [8, 9, 17, 19], as well as the possibility for providing new a priori predictions for new data, which is not the case for probabilistic database programming methods.

Acknowledgements This research was based upon the work supported by NSF SES-1560501 and NSF SES-1560554.

References

1. Abramsky, S. (2013). Relational databases and Bells theorem. In V. Tannen, L. Wong, L. Libkin, W. Fan, T. Wang-Chiew, & M. Fourman (Eds.), *In search of elegance in the theory and practice of computation* (pp. 13–35). Berlin: Springer.
2. Aerts, D., & Aerts, S. (1995). Applications of quantum statistics in psychological studies of decision processes. *Foundations of Science, 1*(1), 85–97.

3. Agresti, A., & Katera, M. (2011). *Categorical data analysis*. Berlin: Springer.
4. Atmanspacher, H., Römer, H., & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics, 32*(3), 379–406.
5. Bordley, R. F. (1998). Quantum mechanical and human violations of compound probability principles: Toward a generalized Heisenberg uncertainty principle. *Operations Research, 46*(6), 923–926.
6. Bruza, P. D., Kitto, K., Ramm, B. J., & Sitbon, L. (2015). A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology, 67*, 26–38.
7. Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge: Cambridge University Press.
8. Busemeyer, J. R., & Wang, Z. (2018). Data fusion using Hilbert space multi-dimensional models. *Theoretical Computer Science, 752*, 41–55.
9. Busemeyer, J. R. & Wang, Z. (2018). Hilbert space multidimensional theory. *Psychological Review, 125*(4), 572–591.
10. Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. New York, NY: Cambridge University Press.
11. Dzhafarov, E., & Kujala, J. V. (2012). Selectivity in probabilistic causality: Where psychology runs into quantum physics. *Journal of Mathematical Psychology, 56*, 54–63.
12. Gleason, A. M. (1957). Measures on the closed subspaces of a Hilbert space. *Journal of Mathematical Mechanics, 6*, 885–893.
13. Gudder, S. P. (1988). *Quantum probability*. Boston, MA: Academic Press.
14. Khrennikov, A. (1999). Classical and quantum mechanics on information spaces with applications to cognitive, psychological, social, and anomalous phenomena. *Foundations of Physics, 29*(7), 1065–1098.
15. Khrennikov, A. Y. (2010). *Ubiquitous quantum structure: From psychology to finance*. Berlin: Springer.
16. Melucci, M. (2015). *Introduction to information retrieval and quantum mechanics*. Berlin: Springer.
17. Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review, 120*(3), 679–696
18. van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge: Cambridge University Press.
19. Wang, Z., & Busemeyer, J. R. (2016). Comparing quantum versus Markov random walk models of judgments measured by rating scales. *Philosophical Transactions of the Royal Society, A, 374*, 20150098.

Basics of Quantum Theory for Quantum-Like Modeling Information Retrieval

Andrei Khrennikov

Abstract This chapter contains a brief introduction to the mathematical formalism and axiomatics of quantum mechanics (QM). Recently quantum mathematics and methodology started to be widely used for modeling decision making for humans and AI-systems, including quantum-like modeling information retrieval. Experts in such areas do not go deeply into the details of quantum theory. Moreover, typically such consumers of quantum theory do not use all its components. Quantum measurement theory is the most useful for application, including information retrieval. The main issue is the quantum treatment of incompatible observables represented mathematically by noncommuting Hermitian operators. At the level of statistical data incompatibility is represented as interference of probabilities, in the form of modification of the formula of total probability by adding the interference term.

1 Introduction

Recently the mathematical formalism and methodology of QM, especially theory of quantum measurement, started to be widely applied outside of physics[1]: to cognition, psychology, economics, finances, decision making, AI, game theory, and information retrieval (for the latter, see, e.g., [24, 51–53, 56, 58, 59] and the chapters in this book). This chapter contains a brief introduction to the mathematical formalism and axiomatics of QM. It is oriented to non-physicists. Since *QM is a statistical theory* it is natural to start with the classical probability

[1] See, for example, [1–10, 15, 27, 28, 32, 38–41, 43–48, 55, 57].

A. Khrennikov (✉)
Linnaeus University, International Center for Mathematical Modeling in Physics and Cognitive Sciences, Växjö, Sweden
e-mail: andrei.khrennikov@lnu.se

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_4

model (Kolmogorov [49]). Then we present basics of theory of Hilbert spaces and Hermitian operators, representation of pure and mixed states by normalized vectors and density operators. This introduction is sufficient to formulate the axiomatics of QM in the form of five postulates. The projection postulate (the most questionable postulate of QM) is presented in a separate section. We distinguish sharply the cases of quantum observables represented by Hermitian operators with nondegenerate and degenerate spectra, the von Neumann's and Lüders' forms of the projection postulate. The axiomatics is completed by a short section on the main interpretations of QM. The projection postulate (in Lüders' form) plays the crucial role in the definition of quantum conditional (transition) probability. By operating with the latter we consider interference of probabilities for two incompatible observables, as a modification of the formula of total probability by adding the interference term. This viewpoint to interference of probabilities was elaborated in a series of works of Khrennikov (see, e.g., [29–38]).

Since classical probability theory is based on the Boolean algebra of events a violation of the law of total probability can be treated as the probabilistic sign of a violation of the laws of the Boolean logics. From this viewpoint, quantum theory can be considered as representing a new kind of logic, the so-called quantum logic. The latter is also briefly presented in a separate section.

We continue this review with a new portion of "quantum mathematics," namely the notion of the tensor product of Hilbert spaces and the tensor product of operators. After the section on Dirac's notation with ket- and bra-vector, we discuss briefly the notion of qubit and entanglement of a few qubits. This chapter is finished with the presentation of the detailed analysis of the probabilistic structure of the two-slit experiment, as a bunch of different experimental contexts. This contextual structure leads to a violation of the law of total probability and the non-Kolmogorovean probabilistic structure of this experiment.

We hope that this chapter would be interesting for newcomers to quantum-like modeling. May be even experts can find something useful, say the treatment of interference of probabilities as a violation of the law of total probability. In any event, this chapter can serve as the mathematical and foundational introduction to other chapters of this book devoted to the concrete applications.

2 Kolmogorov's Axiomatics of Classical Probability

The main aim of QM is to provide probabilistic predictions on the results of measurements. Moreover, statistics of the measurements of a single quantum observable can be described by classical probability theory. In this section we shall present an elementary introduction to this theory.

We remark that classical probability theory is coupled to experiment in the following way:

- Experimental contexts (system's state preparations) are represented by probabilities.
- Observables are represented by random variables.

In principle, we can call probability a state and this is the direct analog of the quantum state (Sect. 6.1, the ensemble interpretation). However, we have to remember that the word "state" has the meaning "statistical state," the state of an ensemble of systems prepared for measurement.

The *Kolmogorov probability space* [49, 50] is any triple

$$(\Omega, \mathcal{F}, \mathbf{P}),$$

where Ω is a set of any origin and $\mathcal{F}$ is a σ-algebra of its subsets, $\mathbf{P}$ is a probability measure on $\mathcal{F}$. The set Ω represents random parameters of the model. Kolmogorov called elements of Ω *elementary events*. This terminology is standard in mathematical literature. Sets of elementary events are regarded as *events*. The key point of Kolmogorov's axiomatization of probability theory is that not any subset of Ω can be treated as an event. For any stochastic model, the system of events $\mathcal{F}$ is selected from the very beginning. The key mathematical point is that $\mathcal{F}$ has to be a *σ-algebra.* (Otherwise it would be very difficult to develop a proper notion of integral. And the latter is needed to define average of a random variable.)

We remind that a σ-algebra is a system of sets which contains Ω and empty set, it is closed with respect to the operations of countable union and intersection and to the operation of taking the complement of a set. For example, the collection of all subsets of Ω is a σ-algebra. This σ-algebra is used in the case of finite or countable sets:

$$\Omega = \{\omega_1, \ldots, \omega_n, \ldots\}. \tag{1}$$

However, for "continuous sets," e.g., $\Omega = [a, b] \subset \mathbf{R}$, the collection of all possible subsets is too large to have applications. Typically it is impossible to describe a σ-algebra in the direct terms. To define a σ-algebra, one starts with a simple system of subsets of Ω and then consider the σ-algebra which is generated from this simple system with the aid of aforementioned operations. In particular, one of the most important for applications σ-algebras, the so-called *Borel σ-algebra*, is constructed in this way by staring with the system consisting of all open and closed intervals of the real line. In a metric space (in particular, in a Hilbert space), the Borel σ-algebra is constructed by starting with the system of all open and closed balls.

Finally, we remark that in American literature the term σ-field is typically used, instead of σ-algebra.

The probability is defined as a *measure*, i.e., a map from $\mathcal{F}$ to nonnegative real numbers which is σ-additive:

$$\mathbf{P}(\cup_j A_j) = \sum_j \mathbf{P}(A_j), \tag{2}$$

where $A_j \in \mathcal{F}$ and $A_i \cap A_j = \emptyset, i \neq j$. The probability measure is always normalized by one:

$$\mathbf{P}(\Omega) = 1. \tag{3}$$

In the case of a discrete probability space, see (1), the probability measures have the form

$$\mathbf{P}(A) = \sum_{\omega_j \in A} p_j, \ \ p_j = \mathbf{P}(\{\omega_j\}).$$

In fact, any finite measure μ, i.e., $\mu(\Omega) < \infty$, can be transformed into the probability measure by normalization:

$$\mathbf{P}(A) = \frac{\mu(A)}{\mu(\Omega)}, A \in \mathcal{F}. \tag{4}$$

A (real) random variable is a map $\xi : \Omega \to \mathbf{R}$ which is measurable with respect to the Borel σ-algebra $\mathcal{B}$ of $\mathbf{R}$ and the σ-algebra $\mathcal{F}$ of Ω. The latter means that, for any set $B \in \mathcal{B}$, its preimage $\xi^{-1}(B) = \{\omega \in \Omega : \xi(\omega) \in B\}$ belongs to $\mathcal{F}$. This condition provides the possibility to assign the probability to the events of the type "values of ξ belong to a (Borel) subset of the real line." The probability distribution of ξ is defined as

$$\mathbf{P}_\xi(B) = \mathbf{P}(\xi^{-1}(B)). \tag{5}$$

In the same way we define the real (and complex) vector-valued random variables, $\xi : \Omega \to \mathbf{R}^n$ and $\xi : \Omega \to \mathbf{C}^n$.

Let $\xi_1, \ldots, \xi_k$ be real-valued random variables. Their join probability distribution $\mathbf{P}_{\xi_1,\ldots,\xi_k}$ is defined as the probability distribution of the vector-valued random variable $\xi = (\xi_1, \ldots, \xi_k)$. To determine this probability measure, it is sufficient to define probabilities

$$\mathbf{P}_{\xi_1,\ldots,\xi_k}(\Gamma_1 \times \cdots \times \Gamma_k) = \mathbf{P}(\omega \in \Omega : \xi_1(\omega) \in \Gamma_1, \ldots., \xi_k(\omega) \in \Gamma_k),$$

where Γ_j, $j = 1, \ldots, k$, are intervals (open, closed, half-open) of the real line.

Suppose now that random variables $\xi_1, \ldots, \xi_k$ represent observables $a_1, \ldots, a_k$. For any point $\omega \in \Omega$, the values of the vector ξ composed of these random variables are well defined $\xi(\omega) = (\xi_1(\omega), \ldots, \xi_k(\omega))$. This vector represents a *joint measurement* of the observables and $\mathbf{P}_{\xi_1,\ldots,\xi_k}$ represents the probability distribution for the outcomes of these jointly measured observables. Thus classical probability theory is applicable for jointly measurable observables, compatible observables in the terminology of QM (Sect. 5).

A random variable is called discrete if its image consists of finite or countable number of points, $\xi = \alpha_1, \ldots, \alpha_n, \ldots$. In this case its probability distribution has the form

$$\mathbf{P}(B) = \sum_{\alpha_j \in B} P_{\alpha_j}, \;\; P_{\alpha_j} = \mathbf{P}(\omega \in \Omega : \xi(\omega) = \alpha_j). \tag{6}$$

The mean value (average) of a real-valued random variable is defined as its integral (the Lebesgue integral)

$$E\xi = \int_{\Omega} \xi(\omega) dP(\omega). \tag{7}$$

For a discrete random variable, its mean value has the simple representation:

$$E\xi = \sum_{\alpha_j \in B} \alpha_j P_{\alpha_j}. \tag{8}$$

In the Kolmogorov model the conditional probability is *defined* by the *Bayes formula*

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(B \cap A)}{\mathbf{P}(A)}, \;\; \mathbf{P}(A) > 0. \tag{9}$$

We stress that other axioms are independent of this definition.

We also present the *formula of total probability* (FTP) which is a simple consequence of the Bayes formula. Consider the pair, a and b, of discrete random variables. Then

$$\mathbf{P}(b = \beta) = \sum_{\alpha} \mathbf{P}(a = \alpha)\mathbf{P}(b = \beta | a = \alpha). \tag{10}$$

Thus the b-probability distribution can be calculated from the a-probability distribution and the conditional probabilities $\mathbf{P}(b = \beta | a = \alpha)$. These conditional probabilities are known as *transition probabilities.*

This formula plays the crucial role in Bayesian inference. It is applicable to the plenty of phenomena, in insurance, finances, economics, engineering, biology, AI, game theory, decision making, and programming. However, as was shown by the author of this review, in quantum domain FTP is violated and it is perturbed by the so-called interference term. Recently it was shown that even data collected in cognitive science, psychology, game theory, and decision making can violate classical FTP [1–10, 18, 38–41, 44–48].

3 Quantum Mathematics

We present the basic mathematical structures of QM and couple them to quantum physics.

3.1 Hermitian Operators in Hilbert Space

We recall the definition of a complex Hilbert space. Denote it by H. This is a complex linear space endowed with a scalar product (a positive-definite nondegenerate Hermitian form) which is complete with respect to the norm corresponding to the scalar product, $\langle\cdot|\cdot\rangle$. The norm is defined as

$$\|\phi\| = \sqrt{\langle\phi|\phi\rangle}.$$

In the finite-dimensional case the norm and, hence, completeness are of no use. Thus those who have no idea about functional analysis (but know essentials of linear algebra) can treat H simply as a finite-dimensional complex linear space with the scalar product.

For a complex number $z = x + iy$, $x, y \in \mathbf{R}$, its conjugate is denoted by $\bar{z}$, here $\bar{z} = x - iy$. The absolute value of z is given by $|z|^2 = z\bar{z} = x^2 + y^2$.

For reader's convenience, we recall that the scalar product is a function from the Cartesian product $H \times H$ to the field of complex numbers $\mathbb{C}$, $\psi_1, \psi_2 \to \langle\psi_1|\psi_2\rangle$, having the following properties:

1. Positive-definiteness: $\langle\psi|\psi\rangle \geq 0$ with $\langle\psi, \psi\rangle = 0$ if and only if $\psi = 0$.
2. Conjugate symmetry: $\langle\psi_1|\psi_2\rangle = \overline{\langle\psi_2|\psi_1\rangle}$
3. Linearity with respect to the second argument[2]: $\langle\phi|k_1\psi_1 + k_2\psi_2\rangle = k_1\langle\phi|\psi_1\rangle + k_2\langle\phi|\psi_2\rangle$, where k_1, k_2 are the complex numbers.

A reader who does not feel comfortable in the abstract framework of functional analysis can simply proceed with the Hilbert space $H = \mathbf{C}^n$, where $\mathbf{C}$ is the set of complex numbers, and the scalar product

$$\langle u|v\rangle = \sum_i u_i \bar{v}_i, u = (u_1, \ldots, u_n), v = (v_1, \ldots, v_n). \tag{11}$$

In this case the above properties of a scalar product can be easily derived from (11). Instead of linear operators, one can consider matrices.

[2]In mathematical texts one typically considers linearity with respect to the first argument. Thus a mathematician has to pay attention to this difference.

We also recall a few basic notions of theory of linear operators in complex Hilbert space. A map $\widehat{a} : H \to H$ is called a linear operator, if it maps linear combination of vectors into linear combination of their images:

$$\widehat{a}(\lambda_1 \psi_1 + \lambda_2 \psi_2) = \lambda_1 \widehat{a} \psi_1 + \lambda_2 \widehat{a} \psi_2,$$

where $\lambda_j \in \mathbf{C}$, $\psi_j \in H$, $j = 1, 2$.

For a linear operator $\widehat{a}$ the symbol $\widehat{a}^*$ denotes its *adjoint operator* which is defined by the equality

$$\langle \widehat{a} \psi_1 | \psi_2 \rangle = \langle \psi_1 | \widehat{a}^* \psi_2 \rangle. \tag{12}$$

Let us select in H some orthonormal basis (e_i), i.e., $\langle e_i | e_j \rangle = \delta_{ij}$. By denoting the matrix elements of the operators $\widehat{a}$ and $\widehat{a}^*$ as a_{ij} and a^*_{ij}, respectively, we rewrite the definition (12) in terms of the matrix elements:

$$a^*_{ij} = \bar{a}_{ji}.$$

A linear operator $\widehat{a}$ is called *Hermitian* if it coincides with its adjoint operator:

$$\widehat{a} = \widehat{a}^\star.$$

If an orthonormal basis in H is fixed, (e_i), and $\widehat{a}$ is represented by its matrix, $A = (a_{ij})$, where $a_{ij} = \langle \widehat{a} e_i | e_j \rangle$, then it is Hermitian if and only if

$$\bar{a}_{ij} = a_{ji}.$$

We remark that, for a Hermitian operator, all its eigenvalues are real. In fact, this was one of the main reasons to represent quantum observables by Hermitian operators. In the quantum formalism, the spectrum of a linear operator (the set of eigenvalues while we are in the finite-dimensional case) coincides with the set of possibly observable values (Sect. 4, **Postulate 3**). We also recall that eigenvectors of Hermitian operators corresponding to different eigenvalues are orthogonal. This property of Hermitian operators plays some role in justification of the projection postulate of QM, see Sect. 5.1.

A linear operator $\widehat{a}$ is *positive-semidefinite* if, for any $\phi \in H$,

$$\langle \widehat{a} \phi | \phi \rangle \geq 0.$$

This is equivalent to positive-semidefiniteness of its matrix.

For a linear operator $\widehat{a}$, its trace is defined as the sum of diagonal elements of its matrix in any orthonormal basis:

$$\mathrm{Tr}\, \widehat{a} = \sum_i a_{ii} = \sum_i \langle \widehat{a} e_i | e_i \rangle,$$

i.e., this quantity does not depend on a basis.

Let L be a subspace of H. The orthogonal projector $P : H \to L$ onto this subspace is a Hermitian, idempotent (i.e., coinciding with its square), and positive-semidefinite operator[3]:

(a) $P^* = P$;
(b) $P^2 = P$;
(c) $P \geq 0$.

Here (c) is a consequence of (a) and (b). Moreover, an arbitrary linear operator satisfying (a) and (b) is an orthogonal projector—onto the subspace PH.

3.2 Pure and Mixed States: Normalized Vectors and Density Operators

Pure quantum states are represented by normalized vectors, $\psi \in H : \|\psi\| = 1$. Two colinear vectors,

$$\psi' = \lambda\psi, \lambda \in \mathbf{C}, |\lambda| = 1, \tag{13}$$

represent the same pure state. Thus, rigorously speaking, a pure state is an equivalence class of vectors having the unit norm: $\psi' \sim \psi$ for vectors coupled by (13). The unit sphere of H is split into disjoint classes—pure states. However, in concrete calculations one typically uses just concrete representatives of equivalent classes, i.e., one works with normalized vectors.

Each pure state can also be represented as the projection operator P_ψ which projects H onto one dimensional subspace based on ψ. For a vector $\phi \in H$,

$$P_\psi \phi = \langle \phi | \psi \rangle \ \psi. \tag{14}$$

The trace of the one dimensional projector P_ψ equals 1:

$$\text{Tr } P_\psi = \langle \psi | \psi \rangle = 1.$$

We summarize the properties of the operator P_ψ representing the pure state ψ. It is

(a) Hermitian,
(b) positive-semidefinite,
(c) trace one,
(d) idempotent.

[3] To simplify formulas, we shall not put the operator-label "hat" in the symbols denoting projectors, i.e., $P \equiv \widehat{P}$.

Moreover, any operator satisfying (a)–(d) represents a pure state. Properties (a) and (d) characterize orthogonal projectors, property (b) is their consequence. Property (c) implies that the projector is one dimensional.

The next step in the development of QM was the extension of the class of quantum states, from pure states represented by one dimensional projectors to states represented by linear operators having the properties (a)–(c). Such operators are called *density operators.* (This nontrivial step of extension of the class of quantum states was based on the efforts of Landau and von Neumann.) The symbol $D(H)$ denotes the space of density operators in the complex Hilbert space H.

One typically distinguishes pure states, as represented by one dimensional projectors, and mixed states, the density operators which cannot be represented by one dimensional projectors. The terminology "mixed" has the following origin: any density operator can be represented as a "mixture" of pure states (ψ_i) :

$$\rho = \sum_i p_i P_{\psi_i}, \;\; p_i \in [0, 1], \sum_i p_i = 1. \tag{15}$$

(To simplify formulas, we shall not put the operator-label "hat" in the symbols denoting density operators, i.e., $\rho \equiv \widehat{\rho}$.) The state is pure if and only if such a mixture is trivial: all p_i, besides one, equal zero. However, by operating with the terminology "mixed state" one has to take into account that the representation in the form (15) is not unique. The same mixed state can be presented as mixtures of different collections of pure states.

Any operator ρ satisfying (a)–(c) is diagonalizable (even in infinite-dimensional Hilbert space), i.e., in some orthonormal basis it is represented as a diagonal matrix, $\rho = \text{diag}(p_j)$, where $p_j \in [0, 1], \sum_j p_j = 1$. Thus it can be represented in the form (15) with mutually orthogonal one dimensional projectors. The property (d) can be used to check whether a state is pure or not.

We point out that *pure states are merely mathematical abstractions; in real experimental situations, it is possible to prepare only mixed states.* The degree of *purity* is defined as

$$\text{purity}(\rho) = \text{Tr}\rho^2.$$

Experimenters are satisfied by getting this quantity near one.

4 Quantum Mechanics: Postulates

We state again that H denotes complex Hilbert space with the scalar product $\langle \cdot, \cdot \rangle$ and the norm $\| \cdot \|$ corresponding to the scalar product.

Postulate 1 (The Mathematical Description of Quantum States) *Quantum (pure) states (wave functions) are represented by normalized vectors ψ (i.e., $\|\psi\|^2 = \langle \psi, \psi \rangle = 1$) of a complex Hilbert space $\mathcal{H}$. Every normalized vector $\psi \in H$ may* *represent a quantum state. If a vector* ψ *corresponding to a state is multiplied by any complex number* c, $|c| = 1$, *the resulting vector will correspond to the same state.*[4]

The physical meaning of "a quantum state" is not defined by this postulate, see Sect. 6.1.

Postulate 2 (The Mathematical Description of Physical Observables) *A physical observable* a *is represented by a Hermitian operator* $\widehat{a}$ *in complex Hilbert space* H. *Different observables are represented by different operators.*

Postulate 3 (Spectral) *For a physical observable* a *which is represented by the Hermitian operator* $\widehat{a}$ *we can predict (together with some probabilities) values* $\lambda \in \mathrm{Spec}(\widehat{a})$.

We restrict our considerations by simplest Hermitian operators which are analogous to discrete random variables in classical probability theory. We recall that a Hermitian operator $\widehat{a}$ has *purely discrete spectrum* if it can be represented as

$$\widehat{a} = \alpha_1 P^a_{\alpha_1} + \cdots + \alpha_m P^a_{\alpha_m} + \cdots, \quad \alpha_m \in \mathbf{R}, \tag{16}$$

where $P^a_{\alpha_m}$ are orthogonal projection operators related to the orthonormal eigenvectors $\{e^a_{km}\}_k$ of $\widehat{a}$ corresponding to the eigenvalues α_m by

$$P^a_{\alpha_m}\psi = \sum_k \langle\psi, e^a_{km}\rangle e^a_{km}, \quad \psi \in H. \tag{17}$$

Here k labels the eigenvectors e^a_{km} which belong to the same eigenvalue α_m of $\widehat{a}$.

Postulate 4 (Born's Rule) *Let a physical observable* a *be represented by a Hermitian operator* $\widehat{a}$ *with purely discrete spectrum. The probability* $\mathbf{P}_\psi(a = \alpha_m)$ *to obtain the eigenvalue* α_m *of* $\widehat{a}$ *for measurement of* a *in a state* ψ *is given by*

$$\mathbf{P}_\psi(a = \alpha_m) = \|P^a_m\psi\|^2. \tag{18}$$

If the operator $\widehat{a}$ has nondegenerate (purely discrete) spectrum, then each α_m is associated with one dimensional subspace. The latter can be fixed by selecting any normalized vector, say e^a_m. In this case orthogonal projectors act simply as

$$P^a_{\alpha_m}\psi = \langle\psi, e^a_m\rangle e^a_m. \tag{19}$$

Formula (18) takes a very simple form

$$\mathbf{P}_\psi(a = \alpha_m) = |\langle\psi, e^a_m\rangle|^2. \tag{20}$$

[4]Thus states are given by elements of the unit sphere of the Hilbert space H.

It is Born's rule in the Hilbert space formalism.

It is important to point out that *if state* ψ *is an eigenstate of operator* $\widehat{a}$ *representing observable* a, i.e., $\widehat{a}\psi = \alpha\psi$, *then the outcome of observable* a *equals* α *with probability one.*

We point out that, for any fixed quantum state ψ, each quantum observable $\widehat{a}$ can be represented as a classical random variable (Sect. 2). In the discrete case the corresponding probability distribution is defined as

$$\mathbf{P}(A) = \sum_{\alpha_m \in A} \mathbf{P}_\psi(a = \alpha_m),$$

where $\mathbf{P}_\psi(a = \alpha_m)$ is given by Born's rule.

Thus each concrete quantum measurement can be described by the classical probability model.

Problems (including deep interpretational issues) arise only when one tries to describe classically data collected for a few incompatible observables (Sect. 5).

By using the Born's rule (18) and the classical probabilistic definition of average (Sect. 2), it is easy to see that the average value of the observable a in the state ψ (belonging to the domain of definition of the corresponding operator $\widehat{a}$) is given by

$$\langle a \rangle_\psi = \langle \widehat{a}\, \psi, \psi \rangle. \tag{21}$$

For example, for an observable represented by an operator with the purely discrete spectrum, we have

$$\langle a \rangle_\psi = \sum_m \alpha_m \mathbf{P}_\psi(a = \alpha_m) = \sum_m \alpha_m \| P_m^a \psi \|^2 = \langle \widehat{a}\, \psi, \psi \rangle.$$

Postulate 5 (Time Evolution of Wave Function) *Let* $\widehat{\mathcal{H}}$ *be the Hamiltonian of a quantum system, i.e., the Hermitian operator corresponding to the energy observable. The time evolution of the wave function* $\psi \in H$ *is described by the Schrödinger equation*

$$i \frac{d}{dt} \psi(t) = \widehat{\mathcal{H}} \psi(t) \tag{22}$$

with the initial condition $\psi(0) = \psi$.

5 Compatible and Incompatible Observables

Two observables a and b are called *compatible* if a measurement procedure for their joint measurement can be designed, i.e., a measurement of the vector observable $d = (a, b)$. In such a case their joint probability distribution is well defined.

In the opposite case, i.e., when such a joint-measurement procedure does not exist, observables are called *incompatible.* The joint probability distribution of incompatible observables has no meaning.

In QM, compatible observables a and b are represented by commuting Hermitian operators $\widehat{a}$ and $\widehat{b}$, i.e., $[\widehat{a}, \widehat{b}] = 0$; consequently, incompatible observables a and b are represented by noncommuting operators, i.e., $[\widehat{a}, \widehat{b}] \neq 0$. Thus in the QM-formalism compatibility–incompatibility is represented as commutativity–noncommutativity.

Postulate 4a (Born's Rule for Joint Measurements) *Let observables a and b be represented by Hermitian operators $\widehat{a}$ and $\widehat{b}$ with purely discrete spectrum. The probability to obtain the eigenvalues α_m and β_k in a joint measurement of a and b in a state ψ—the joint probability distribution—is given by*

$$\mathbf{P}_\psi(a = \alpha_m, b = \beta_k) = \|P_k^b P_m^a \psi\|^2 = \|P_m^a P_k^b \psi\|^2. \tag{23}$$

It is crucial that the spectral projectors of commuting operators commute, so the probability distribution does not depend on the order of the values of observables. This is a classical probability distribution (Sect. 2). Any pair of compatible observables a and b can be represented by random variables: for example, by using the joint probability distribution as the probability measure.

A family of compatible observables $a_1, \ldots, a_n$ is represented by commuting Hermitian operators $\widehat{a}_1, \ldots, \widehat{a}_n$, i.e., $[\widehat{a}_i, \widehat{a}_j] = 0$ for all pairs i, j. The joint probability distribution is given by the natural generalization of rule (23):

$$\mathbf{P}_{a_1,\ldots,a_n;\psi}(\alpha_{1m}, \ldots, \alpha_{nk}) \equiv \mathbf{P}_\psi(a_1 = \alpha_{1m}, \ldots, a_n = \alpha_{nk})$$

$$= \|P_k^{a_n} \ldots .P_m^{a_1}\psi\|^2 = \cdots . = \|P_m^{a_1} \ldots .P_k^{a_n}\psi\|^2, \tag{24}$$

where all possible permutations of projectors can be considered.

Now we point to one distinguishing feature of compatibility of *quantum observables* which is commonly not emphasized. The relation of commutativity of operators is the pairwise relation, it does not involve say triples of operators. Thus, for joint measurability of a group of quantum observables $a_1, \ldots, a_n$, their pairwise joint measurability is sufficient. Thus if we are able to design measurement procedures for all possible pairs, then we are always able to design a joint-measurement procedure for the whole group of quantum observables $a_1, \ldots, a_n$. This is the specialty of quantum observables. In particular, if there exist all pairwise joint probability distributions $\mathbf{P}_{a_i,a_j;\psi}$, then the joint probability $\mathbf{P}_{a_1,\ldots,a_n;\psi}$ is defined as well.

The Born's rule can be generalized to the quantum states represented by density operators (Sect. 9.1, formula (40)).

5.1 Post-Measurement State From the Projection Postulate

The projection postulate is one of the most questionable and debatable postulates of QM. We present it in the separate section to distinguish it from other postulates of QM, Postulates 1–5, which are commonly accepted.

Consider pure state ψ and quantum observable (Hermitian operator) $\widehat{a}$ representing some physical observable a. Suppose that $\widehat{a}$ has nondegenerate spectrum; denote its eigenvalues by $\alpha_1, .., \alpha_m, \ldots$ and the corresponding eigenvectors by $e_1^a, \ldots, e_m^a, \ldots$ (here $\alpha_i \neq \alpha_j, i \neq j$.) This is an orthonormal basis in H. We expand the vector ψ with respect to this basis:

$$\psi = k_1 e_1^a + \cdots + k_m e_m^a + \cdots , \tag{25}$$

where (k_j) are complex numbers such that

$$\|\psi\|^2 = |k_1|^2 + \cdots + |k_m|^2 + \cdots = 1. \tag{26}$$

By using the terminology of linear algebra we say that the pure state ψ is a *superposition* of the pure states e_j. The von Neumann projection postulate describes the post-measurement state and it can be formulated as follows:

Postulate 6VN (Projection Postulate, von Neumann) *Measurement of observable a resulting in output α_i induces reduction of superposition (25) to the basis vector e_i^a.*

The procedure described by the projection postulate can be interpreted in the following way:

Superposition (25) reflects uncertainty in the results of measurements for an observable a. Before measurement a quantum system "does not know how it will answer to the question a." The Born's rule presents potentialities for different answers. Thus a quantum system in the superposition state ψ does not have propensity to any value of a as its objective property. After the measurement the superposition is reduced to the single term in the expansion (25) corresponding to the value of a obtained in the process of measurement.

Consider now an arbitrary quantum observable a with purely discrete spectrum, i.e., $\widehat{a} = \alpha_1 P_{\alpha_1}^a + \cdots + \alpha_m P_{\alpha_m}^a + \cdots$. The Lüders projection postulate describes the post-measurement state and it can be formulated as follows:

Postulate 6L (Projection Postulate, Lüders) *Measurement of observable a resulting in output α_m induces projection of state ψ on state*

$$\psi_{\alpha_m} = \frac{P_{\alpha_m}^a \psi}{\|P_{\alpha_m}^a \psi\|}.$$

In contrast to the majority of books on quantum theory, we sharply distinguish the cases of quantum observables with nondegenerate and degenerate spectra. von

Neumann formulated Postulate 6VN only for observables with nondegenerate spectra. Lüders "generalized" von Neumann's postulate to the case of observables with degenerate spectra. However, for such observables, von Neumann formulated [60] a postulate which is different from Postulate 6L. The post-measurement state need not be again a pure state.

We remark that Postulate 6L can be applied even to quantum states which are represented by density operators (Sect. 9.1, formula (41)).

6 Interpretations of Quantum Mechanics

Now we are going to discuss one of the most important and complicated issues of quantum foundations, the problem of an interpretation of a quantum state. There were elaborated numerous interpretations which can differ crucially from each other. This huge diversity of interpretations is a sign of the deep crises in quantum foundations.

In this section, we briefly discuss a few basic interpretations. Then in Sect. 9.1 we shall consider the Växjö (realist ensemble contextual) interpretation. Its presentation needs additional mathematical formulas. Therefore we placed it into a separate section.

6.1 Ensemble and Individual Interpretations

The Ensemble Interpretation *A (pure) quantum state provides a description of certain statistical properties of an ensemble of similarly prepared quantum systems.*

This interpretation is upheld, for example, by Einstein, Popper, Blokhintsev, Margenau, Ballentine, Klyshko, and recent years by, e.g., De Muynck, De Baere, Holevo, Santos, Khrennikov, Nieuwenhuizen, Adenier, Groessing, and many others.

The Copenhagen Interpretation *A (pure) quantum state provides the complete description of an individual quantum system.*

This interpretation was supported by a great variety of members, from Schrödinger's original attempt to identify the electron with a wave function solution of his equation to the several versions of the Copenhagen interpretation [12–14, 53, 54] (for example, Heisenberg, Bohr, Pauli, Dirac, von Neumann, Landau, Fock, and recent years by, e.g., Greenberger, Mermin, Lahti, Peres, Summhammer). Nowadays the individual interpretation is extremely popular, especially in quantum information and computing.

Instead of Einstein's terminology "*ensemble interpretation*," Ballentine [7–9] used the terminology *"statistical interpretation."* However, Ballentine's terminology is rather misleading, because the term "statistical interpretation" was also used by von Neumann [60] for individual randomness! For him "statistical interpretation" had the meaning which is totally different from the Ballentine's "ensemble-statistical interpretation." J. von Neumann wanted to emphasize the difference between deterministic (Newtonian) classical mechanics in that the state of a system is determined by values of two observables (position and momentum) and quantum mechanics in that the state is determined not by values of observables, but by probabilities. We shall follow Albert Einstein and use the terminology *ensemble interpretation.*

We remark that following von Neumann [60] the supporters of the individual interpretation believe in *irreducible quantum randomness,* i.e., that the behavior of an individual quantum system is irreducibly random. Why does it behave in such a way? Because it is quantum, so it can behave so unusually. Nowadays this von Neumann's claim is used to justify superiority of the quantum technology over the classical technology. For example, superiority of quantum random generators.

6.2 Information Interpretations

The quantum information revolution generated a variety of information interpretations of QM (see, for example, [16, 17, 19, 20]). By these interpretations the quantum formalism describes special way of information processing, more general than the classical information processing. Roughly speaking one can forget about physics and work solely with probability, entropy, and information. Quantum Bayesianism (QBism) [25, 26] can be considered as one of such information, in its extreme form: the quantum formalism describes very general scheme of assignments of subjective probabilities to possible outcomes of experiments, assignment by human agents.

7 Quantum Conditional (Transition) Probability

In the classical Kolmogorov probabilistic model (Sect. 2), besides probabilities one operates with the conditional probabilities defined by the *Bayes formula* (see Sect. 2, formula (9)). The Born's postulate defining quantum probability should also be completed by a definition of the conditional probability. We have remarked that, for one concrete observable, the probability given by Born's rule can be treated classically. However, the definition of the conditional probability involves two observables. Such situations cannot be treated classically (at least straightforwardly, cf. Sect. 2). Thus conditional probability is really a quantum probability.

Let physical observables a and b be represented by Hermitian operators with purely discrete (may be degenerate) spectra:

$$\widehat{a} = \sum_m \alpha_m P^a_{\alpha_m}, \ \widehat{b} = \sum_m \beta_m P^b_{\beta_m}. \tag{27}$$

Let ψ be a pure state and let $P^a_{\alpha_k}\psi \neq 0$. Then the probability to get the value $b = \beta_m$ under the condition that the value $a = \alpha_k$ was observed in the preceding measurement of the observable a on the state ψ is given by the formula

$$\mathbf{P}_\psi(b = \beta_m | a = \alpha_k) \equiv \frac{\| P^b_{\beta_m} \; P^a_{\alpha_k} \psi \|^2}{\| P^a_{\alpha_k} \; \psi \|^2}. \tag{28}$$

One can motivate this definition by appealing to the projection postulate (Lüders' version). After the a-measurement with output $a = \alpha_k$ initially prepared state ψ is projected onto the state

$$\psi_{\alpha_k} = \frac{P^a_{\alpha_k} \psi}{\| P^a_{\alpha_k} \; \psi \|}.$$

Then one applies Born's rule to the b-measurement for this state.

Let the operator $\widehat{a}$ has nondegenerate spectrum, i.e., for any eigenvalue α the corresponding eigenspace (i.e., generated by eigenvectors with $\widehat{a}\psi = \alpha\psi$) is one dimensional. We can write

$$\mathbf{P}_\psi(b = \beta_m | a = \alpha_k) = \| P^b_{\beta_m} e^a_k \|^2 \tag{29}$$

(here $\widehat{a} e^a_k = \alpha_k e^a_k$). Thus the conditional probability in this case does not depend on the original state ψ. We can say that the memory about the original state was destroyed. If also the operator $\widehat{b}$ has nondegenerate spectrum, then we have: $\mathbf{P}_\psi(b = \beta_m | a = \alpha_k) = |\langle e^b_m, e^a_k \rangle|^2$ and $\mathbf{P}_\psi(a = \alpha_k | b = \beta_m) = |\langle e^a_k, e^b_m \rangle|^2$. By using symmetry of the scalar product we obtain the following result:

Let both operators $\widehat{a}$ and $\widehat{b}$ have purely discrete nondegenerate spectra and let $P^a_k \psi \neq 0$ and $P^b_m \psi \neq 0$. Then conditional probability is symmetric and it does not depend on the original state ψ :

$$\mathbf{P}_\psi(b = \beta_m | a = \alpha_k) = \mathbf{P}_\psi(a = \alpha_k | b = \beta_m) = |\langle e^b_m, e^a_k \rangle|^2. \tag{30}$$

8 Observables with Nondegenerate Spectra: Double-Stochasticity of the Matrix of Transition Probabilities

We remark that classical (Kolmogorov–Bayes) conditional probability is not symmetric, besides very special situations. Thus *QM is described by a very specific probabilistic model.*

Consider two nondegenerate observables. Set $p_{\beta|\alpha} = \mathbf{P}(b = \beta | a = \alpha)$. The matrix of transition probabilities $\mathbf{P}^{b|a} = (p_{\beta|\alpha})$ is not only *stochastic,* i.e.,

$$\sum_\beta p_{\beta|\alpha} = 1$$

but it is even *doubly stochastic*:

$$\sum_{\alpha} p_{\beta|\alpha} = \sum_{\alpha} |\langle e^b_\beta, e^a_\alpha \rangle|^2 = \langle e^b_\beta, e^b_\beta \rangle = 1.$$

In Kolmogorov's model, stochasticity is the general property of matrices of transition probabilities. However, in general classical matrices of transition probabilities are not doubly stochastic. Hence, double stochasticity is a very specific property of quantum probability.

We remark that statistical data collected outside quantum physics, e.g., in decision making by humans and psychology, violates the quantum law of double stochasticity [38]. Such data cannot be mathematically represented with the aid of Hermitian operators with nondegenerate spectra. One has to consider either Hermitian operators with degenerate spectra or positive operator valued measures (POVMs).

9 Formula of Total Probability with the Interference Term

We shall show that the quantum probabilistic calculus violates the conventional FTP (10), one of the basic laws of classical probability theory. In this section, we proceed in the abstract setting by operating with two abstract incompatible observables. The concrete realization of this setting for the two-slit experiment demonstrating interference of probabilities in QM will be presented in Sect. 16 which is closely related to Feynman's claim [22, 23] on the nonclassical probabilistic structure of this experiment.

Let $H_2 = \mathbf{C} \times \mathbf{C}$ be the two dimensional complex Hilbert space and let $\psi \in H_2$ be a quantum state. Let us consider two dichotomous observables $b = \beta_1, \beta_2$ and $a = \alpha_1, \alpha_2$ represented by Hermitian operators $\widehat{b}$ and $\widehat{a}$, respectively (one may consider simply Hermitian matrices). Let $e^b = \{e^b_\beta\}$ and $e^a = \{e^a_\alpha\}$ be two orthonormal bases consisting of eigenvectors of the operators. The state ψ can be represented in the two ways

$$\psi = c_1 e^a_1 + c_2 e^a_2,\ c_\alpha = \langle \psi, e^a_\alpha \rangle; \tag{31}$$

$$\psi = d_1 e^b_1 + d_2 e^b_2,\ d_\beta = \langle \psi, e^b_\beta \rangle. \tag{32}$$

By Postulate 4 we have

$$\mathbf{P}(a = \alpha) \equiv \mathbf{P}_\psi(a = \alpha) = |c_\alpha|^2. \tag{33}$$

$$\mathbf{P}(b = \beta) \equiv \mathbf{P}_\psi(b = \beta) = |d_\beta|^2. \tag{34}$$

The possibility to expand one basis with respect to another basis induces connection between the probabilities $\mathbf{P}(a = \alpha)$ and $\mathbf{P}(b = \beta)$. Let us expand the vectors e^a_α with respect to the basis e^b

$$e^a_1 = u_{11} e^b_1 + u_{12} e^b_2 \tag{35}$$

$$e^a_2 = u_{21} e^b_1 + u_{22} e^b_2, \tag{36}$$

where $u_{\alpha\beta} = \langle e^a_\alpha, e^b_\beta \rangle$. Thus $d_1 = c_1 u_{11} + c_2 u_{21},\ d_2 = c_1 u_{12} + c_1 u_{22}$. We obtain the *quantum rule* for transformation of probabilities:

$$\mathbf{P}(b = \beta) = |c_1 u_{1\beta} + c_2 u_{2\beta}|^2. \tag{37}$$

On the other hand, by the definition of quantum conditional probability, see (28), we obtain

$$\mathbf{P}(b = \beta | a = \alpha) \equiv \mathbf{P}_\psi(b = \beta | a = \alpha) = |\langle e^a_\alpha, e^b_\beta \rangle|^2. \tag{38}$$

By combining (33), (34) and (37), (38) we obtain the *quantum formula of total probability—the formula of the interference of probabilities*:

$$\mathbf{P}(b = \beta) = \sum_\alpha \mathbf{P}(a = \alpha) \mathbf{P}(b = \beta | a = \alpha) \tag{39}$$

$$+2\cos\theta \sqrt{\mathbf{P}(a = \alpha_1)\mathbf{P}(b = \beta | a = \alpha_1)\mathbf{P}(a = \alpha_2)\mathbf{P}(b = \beta | a = \alpha_2)}.$$

In general $\cos\theta \neq 0$. Thus the quantum FTP does not coincide with the classical FTP (10) which is based on the Bayes' formula.

We presented the derivation of the quantum FTP only for observables given by Hermitian operators acting in the two dimensional Hilbert space and for pure states. In Sect. 9.1, we give (without proving) the formula for spaces of an arbitrary dimension and states represented by density operators (see [42] for quantum FTP for observables represented by POVMs).

9.1 Växjö (Realist Ensemble Contextual) Interpretation of Quantum Mechanics

The Växjö interpretation [33] is the *realist ensemble contextual interpretation* of QM. Thus, in contrast to Copenhagenists or QBists, by the Växjö interpretation QM is not complete and it can be emergent from a subquantum model. This interpretation is the ensemble interpretation This interpretation is contextual, i.e., experimental contexts have to be taken into account really seriously.

By the Växjö interpretation *the probabilistic part of QM is a special mathematical formalism to work with contextual probabilities for families of contexts*, which are, in general, incompatible. Of course, the quantum probabilistic formalism is not the only possible formalism to operate with contextual probabilities.

The main distinguishing feature of the formalism of quantum probability is its complex Hilbert space representation and the Born's rule. *All quantum contexts can be unified with the aid of a quantum state* ψ (wave function, complex probability amplitude). A state represents only a part of context, the second part is given by an observable a. Thus the quantum probability model is not just a collection of Kolmogorov probability spaces. These spaces are coupled by quantum states.

Each theory of probability has two main purposes: descriptive and predictive. In classical probability theory its predictive machinery is based on *Bayesian inference* and, in particular, FTP (Sect. 2, formula (10)).

Can the probabilistic formalism of QM be treated as a generalization of Bayesian inference?

My viewpoint is that the quantum FTP with the interference term (Sect. 9, formula (39)) is, in fact, a modified rule for the probability update. QM provides the following inference machinery. There are given a mixed state represented by density operator ρ and two quantum observables a and b represented mathematically by Hermitian operators $\widehat{a}$ and $\widehat{b}$ with purely discrete spectra. The first measurement of a can be treated as collection of information about the state ρ. The result $a = \alpha_i$ appears with the probability

$$p^a(\alpha_i) = \mathrm{Tr} P_i^a \rho. \tag{40}$$

This is generalization of the Born's rule to mixed states.

Postulate 6L (the projection postulate in the Lüders' form) can be extended to mixed states. Initial state ρ is transferred to the state

$$\rho_{a_i} = \frac{P_i^a \rho P_i^a}{\mathrm{Tr} P_i^a \rho P_i^a}. \tag{41}$$

Then, for each state ρ_{a_i}, we perform measurement of b and obtain probabilities

$$p(\beta_j|\alpha_i) = \mathrm{Tr} P_j^b \rho_{a_i}. \tag{42}$$

These are quantum conditional (transition) probabilities for the initial state given by a density operator (generalization of the formalism of Sect. 7).

We now recall the general form of the quantum FTP [42]:

$$p(b = \beta) = \sum_k p(b = \beta|a = \alpha_k) p(a = \alpha_k) \tag{43}$$

$$+2 \sum_{k<m} \cos \phi_{j;k,m} \sqrt{p(b = \beta|a = \alpha_k) p(a = \alpha_k) p(b = \beta|\alpha_m) p(a = \alpha_m)}.$$

Thus we can predict the probability of the result β_j for the b-observable on the basis of the probabilities for the results α_i for the a-observable and conditional probabilities. Of course, the main nonclassical feature of this probability update rule is the presence of phase angles. In the case of dichotomous observables of the von Neumann–Lüders type the phase angles ϕ_j can be expressed in terms of probabilities.

10 Quantum Logic

von Neumann and Birkhoff [11, 61] suggested to represent *events* (propositions) by orthogonal projectors in complex Hilbert space H.

For an orthogonal projector P, we set $H_P = P(H)$, its image, and vice versa, for subspace L of H, the corresponding orthogonal projector is denoted by the symbol P_L.

The set of orthogonal projectors is a *lattice* with the order structure: $P \leq Q$ iff $H_P \subset H_Q$ or equivalently, for any $\psi \in H$, $\langle\psi|P\psi\rangle \leq \langle\psi|Q\psi\rangle$.

We recall that the lattice of projectors is endowed with operations "and" ($\wedge$) and "or" ($\vee$). For two projectors P_1, P_2, the projector $R = P_1 \wedge P_2$ is defined as the projector onto the subspace $H_R = H_{P_1} \cap H_{P_2}$ and the projector $S = P_1 \vee P_2$ is defined as the projector onto the subspace H_R defined as the minimal linear subspace containing the set-theoretic union $H_{P_1} \cup H_{P_2}$ of subspaces H_{P_1}, H_{P_2} : this is the space of all linear combinations of vectors belonging to these subspaces. The operation of negation is defined as the orthogonal complement: $P^{\perp} = \{y \in H : \langle y|x\rangle = 0 \text{ for } \text{ all } x \in H_P\}$.

In the language of subspaces the operation "and" coincides with the usual set-theoretic intersection, but the operations "or" and "not" are nontrivial deformations of the corresponding set-theoretic operations. It is natural to expect that such deformations can induce deviations from classical Boolean logic.

Consider the following simple example. Let H be two dimensional Hilbert space with the orthonormal basis (e_1, e_2) and let $v = (e_1 + e_2)/\sqrt{2}$. Then $P_v \wedge P_{e_1} = 0$ and $P_v \wedge P_{e_2} = 0$, but $P_v \wedge (P_{e_1} \vee P_{e_2}) = P_v$. Hence, for quantum events, in general the distributivity law is violated:

$$P \wedge (P_1 \vee P_2) \neq (P \wedge P_1) \vee (P \wedge P_2). \tag{44}$$

The lattice of orthogonal projectors is called *quantum logic*. It is considered as a (very special) generalization of classical Boolean logic. Any sub-lattice consisting of commuting projectors can be treated as classical Boolean logic.

At the first sight the representation of events by projectors/linear subspaces might look exotic. However, this is simply a prejudice which springs from too common usage of the set-theoretic representation of events (Boolean logic) in the modern classical probability theory. The tradition to represent events by subsets was firmly established by A. N. Kolmogorov in 1933. We remark that before him the basic classical probabilistic models were not of the set-theoretic nature. For example, the main competitor of the Kolmogorov model, the von Mises frequency model, was based on the notion of a collective.

As we have seen, quantum logic relaxes some constraints posed on the operations of classical Boolean logic, in particular, the distributivity constraint. This provides novel possibilities for logically consistent reasoning.

Since human decision makers violate FTP [32, 38]—the basic law of classical probability, it seems that they process information by using nonclassical logic. Quantum logic is one of the possible candidates for logic of human reasoning. However, one has to remember that *in principle other types of nonclassical logic may be useful for mathematical modeling of human decision making.*

11 Space of Square Integrable Functions as a State Space

Although we generally proceed with finite-dimensional Hilbert spaces, it is useful to mention the most important example of infinite-dimensional Hilbert space used in QM. Consider the space of complex valued functions, $\psi : \mathbb{R}^m \to \mathbb{C}$, which are square integrable with respect to the Lebesgue measure on $\mathbb{R}^m$:

$$\|\psi\|^2 = \int_{\mathbb{R}^m} |\psi(x)|^2 dx < \infty. \tag{45}$$

It is denoted by the symbol $L^2(\mathbb{R}^m)$. Here the scalar product is given by

$$\langle\psi_1|\psi_2\rangle = \int_{\mathbb{R}^m} \bar{\psi}_1(x)\psi_2(x)dx.$$

A delicate point is that, for some measurable functions, $\psi : \mathbb{R}^m \to \mathbb{C}$, which are not identically zero, the integral

$$\int_{\mathbb{R}^m} |\psi(x)|^2 dx = 0. \tag{46}$$

We remark that the latter equality implies that $\psi(x) = 0$ a.e. (almost everywhere). Thus the quantity defined by (45) is, in fact, not norm: $\|\psi\| = 0$ does not imply that $\psi = 0$. To define a proper Hilbert space, one has to consider as its elements not simply functions, but classes of equivalent functions, where the equivalence relation is defined as $\psi \sim \phi$ if and only if $\psi(x) = \phi(x)$ a.e. In particular, all functions satisfying (46) are equivalent to the zero-function.

12 Operation of Tensor Product

Let both state spaces be L^2-spaces, the spaces of complex valued square integrable functions: $H_1 = L^2(\mathbb{R}^k)$ and $L^2(\mathbb{R}^m)$.

Take two functions: $\psi \equiv \psi(x)$ belongs to H_1 and $\phi \equiv \phi(y)$ belongs to H_2. By multiplying these functions we obtain the function of two variables $\Psi(x, y) = \psi(x) \times \phi(y)$, where $\times$ denotes the usual point wise product.[5] It is easy to check that this function belongs to the space $H = L^2(\mathbb{R}^{k+m})$. Take now n functions, $\psi_1(x), \ldots, \psi_n(x)$, from H_1 and n functions, $\phi_1(y), \ldots \phi_n(y)$, from H_2 and consider the sum of their pairwise products:

$$\Psi(x, y) = \sum_i \psi_i(x) \times \phi_i(y). \tag{47}$$

This function also belongs to H.

It is possible to show that any function belonging to H can be represented as (47), where the sum is in general infinite. Multiplication of functions is the basic example of the operation of the tensor product. The latter is denoted by the symbol $\otimes$. Thus in the example under consideration $\psi \otimes \phi(x, y) = \psi(x) \times \phi(y)$. The tensor product structure on $H = L^2(\mathbb{R}^{k+m})$ is symbolically denoted as $H = H_1 \otimes H_2$.

Consider now two arbitrary orthonormal bases in spaces $H_k, (e_j^{(k)}), k = 1, 2$. Then functions $(e_{ij} = e_i^{(1)} \otimes e_j^{(2)})$ form an orthonormal basis in H : any $\Psi \in H$ can be represented as

$$\Psi = \sum c_{ij} e_{ij} \equiv \sum c_{ij} e_i^{(1)} \otimes e_j^{(2)}, \tag{48}$$

where

$$\sum |c_{ij}|^2 < \infty. \tag{49}$$

Consider now two arbitrary finite-dimensional Hilbert spaces, H_1, H_2. For each pair of vectors $\psi \in H_1, \phi \in H_2$, we form a new formal entity denoted by $\psi \otimes \phi$. Then we consider the sums $\Psi = \sum_i \psi_i \otimes \phi_i$. On the set of such formal sums we can introduce the linear space structure. (To be mathematically rigorous, we have to constraint this set by some algebraic relations to make the operations of addition and multiplication by complex numbers well defined.) This construction gives us the tensor product $H = H_1 \otimes H_2$. In particular, if we take orthonormal bases in $H_k, (e_j^{(k)}), k = 1, 2$, then $(e_{ij} = e_i^{(1)} \otimes e_j^{(2)})$ form an orthonormal basis in H, any $\Psi \in H$ can be represented as (48) with (49).

[5] Here it is convenient to use this symbol, not just write as $\Psi(x, y) = \psi(x)\phi(y)$.

The latter representation gives the simplest possibility to define the tensor product of two arbitrary (i.e., may be infinite-dimensional) Hilbert spaces as the space of formal series (48) satisfying the condition (49).

Besides the notion of the tensor product of states, we shall also use the notion of the tensor product of operators. Consider two linear operators $\widehat{a}_i : H_i \to H_i, i = 1, 2$. Their tensor product $\widehat{a} \equiv \widehat{a}_1 \otimes \widehat{a}_2 : H \to H$ is defined starting with the tensor products of two vectors:

$$\widehat{a}\ \psi \otimes \phi = (\widehat{a}_1 \psi) \otimes (\widehat{a}_2 \phi).$$

Then it is extended by linearity. By using the coordinate representation (48) the tensor product of operators can be represented as

$$\widehat{a}\ \Psi = \sum c_{ij}\ \widehat{a} e_{ij} \equiv \sum c_{ij} \widehat{a}_1 e_i^{(1)} \otimes \widehat{a}_2 e_j^{(2)}, \tag{50}$$

If operators $\widehat{a}_i, i = 1, 2$, are represented by matrices $(a_{kl}^{(i)})$, with respect to the fixed bases, then the matrix $(a_{kl.nm})$ of the operator $\widehat{a}$ with respect to the tensor product of these bases can be easily calculated.

In the same way one defines the tensor product of Hilbert spaces, $H_1, \ldots, H_n$, denoted by the symbol $H = H_1 \otimes \cdots \otimes H_n$. We start with forming the formal entities $\psi_1 \otimes \cdots \otimes \psi_n$, where $\psi_j \in H_j, j = 1, \ldots, n$. Tensor product space is defined as the set of all sums $\sum_j \psi_{1j} \otimes \cdots \otimes \psi_{nj}$ (which has to be constrained by some algebraic relations, but we omit such details). Take orthonormal bases in $H_k, (e_j^{(k)}), k = 1, \ldots, n$. Then any $\Psi \in H$ can be represented as

$$\Psi = \sum_\alpha c_\alpha e_\alpha \equiv \sum_{\alpha=(j_1 \ldots j_n)} c_{j_1 \ldots j_n} e_{j_1}^{(1)} \otimes \cdots \otimes e_{j_n}^{(n)}, \tag{51}$$

where $\sum_\alpha |c_\alpha|^2 < \infty$.

13 Ket- and Bra-Vectors: Dirac's Symbolism

Dirac's notations [21] are widely used in quantum information theory. Vectors of H are called *ket-vectors*, they are denoted as $|\psi\rangle$. The elements of the dual space H' of H, the space of linear continuous functionals on H, are called *bra-vectors*, they are denoted as $\langle\psi|$.

Originally the expression $\langle\psi|\phi\rangle$ was used by Dirac for the duality form between H' and H, i.e., $\langle\psi|\phi\rangle$ is the result of application of the linear functional $\langle\psi|$ to the vector $|\phi\rangle$. In mathematical notation it can be written as follows. Denote the functional $\langle\psi|$ by f and the vector $|\phi\rangle$ by simply ϕ. Then $\langle\psi|\phi\rangle \equiv f(\phi)$. To simplify the model, later Dirac took the assumption that H is Hilbert space, i.e., the H' can be identified with H. We remark that this assumption is an axiom simplifying the mathematical model of QM. However, in principle Dirac's formalism [21] is applicable for any topological linear space H and its dual space H'; so it is more general than von Neumann's formalism [60] rigidly based on Hilbert space.

Consider an observable a given by the Hermitian operator $\widehat{a}$ with nondegenerate spectrum and restrict our consideration to the case of finite dimensional H. Thus the normalized eigenvectors e_i of A form the orthonormal basis in H. Let $\widehat{a}e_i = \alpha_i e_i$. In Dirac's notation e_i is written as $|\alpha_i\rangle$ and, hence, any pure state can be written as

$$|\psi\rangle = \sum_i c_i|\alpha_i\rangle, \ \sum_i |c_i|^2 = 1. \tag{52}$$

Since the projector onto $|\alpha_i\rangle$ is denoted as $P_{\alpha_i} = |\alpha_i\rangle\langle\alpha_i|$, the operator $\widehat{a}$ can be written as

$$\widehat{a} = \sum_i \alpha_i|\alpha_i\rangle\langle\alpha_i|. \tag{53}$$

Now consider two Hilbert spaces H_1 and H_2 and their tensor product $H = H_1 \otimes H_2$. Let $(|\alpha_i\rangle)$ and $(|\beta_i\rangle)$ be orthonormal bases in H_1 and H_2 corresponding to the eigenvalues of two observables A and B. Then vectors $|\alpha_i\rangle \otimes |\beta_j\rangle$ form the orthonormal basis in H. Typically in physics the sign of the tensor product is omitted and these vectors are written as $|\alpha_i\rangle|\beta_j\rangle$ or even as $|\alpha_i\beta_j\rangle$. Thus any vector $\psi \in H = H_1 \otimes H_2$ can be represented as

$$\psi = \sum_{ij} c_{ij}|\alpha_i\beta_j\rangle, \tag{54}$$

where $c_{ij} \in \mathbf{C}$ (in the infinite-dimensional case these coefficients are constrained by the condition $\sum_{ij} |c_{ij}|^2 < \infty$).

14 Qubit

In particular, in quantum information theory typically qubit states are represented with the aid of observables having the eigenvalues 0, 1. Each qubit space is two dimensional:

$$|\psi\rangle = c_0|0\rangle + c_1|1\rangle, \ |c_0|^2 + |c_1|^2 = 1. \tag{55}$$

A pair of qubits is represented in the tensor product of single qubit spaces, here pure states can be represented as superpositions:

$$|\psi\rangle = c_{00}|00\rangle + c_{01}|01\rangle + c_{10}|10\rangle + c_{11}|00\rangle, \tag{56}$$

where $\sum_{ij} |c_{ij}|^2 = 1$. In the same way the n-qubit state is represented in the tensor product of n one-qubit state spaces (it has the dimension 2^n) :

$$|\psi\rangle = \sum_{x_j=0,1} c_{x_1 \ldots x_n} |x_1 \ldots x_n\rangle, \tag{57}$$

where $\sum_{x_j=0,1} |c_{x_1 \ldots x_n}|^2 = 1$. We remark that the dimension of the n qubit state space grows exponentially with the growth of n. The natural question about possible physical realizations of such multi-dimensional state spaces arises. The answer to it is not completely clear; it depends very much on the used interpretation of the wave function.

15 Entanglement

Consider the tensor product $H = H_1 \otimes H_2 \otimes \cdots \otimes H_n$ of Hilbert spaces $H_k, k = 1, 2, \ldots, n$. The states of the space H can be *separable and non-separable* (entangled). We start by considering pure states. The states from the first class, separable pure states, can be represented in the form:

$$|\psi\rangle = \otimes_{k=1}^{n} |\psi_k\rangle = |\psi_1 \ldots \psi_n\rangle, \tag{58}$$

where $|\psi_k\rangle \in H_k$. The states which cannot be represented in this way are called non-separable or *entangled.* Thus mathematically the notion of entanglement is very simple, it means impossibility of tensor factorization.

For example, let us consider the tensor product of two one-qubit spaces. Select in each of them an orthonormal basis denoted as $|0\rangle, |1\rangle$. The corresponding orthonormal basis in the tensor product has the form $|00\rangle, |01\rangle, |10\rangle, |11\rangle$. Here we used Dirac's notations, see Sect. 13, near the end. Then the so-called Bell's states

$$|\Phi^+\rangle = (|00\rangle + |11\rangle)/\sqrt{2}, \quad |\Phi^-\rangle = (|00\rangle - |11\rangle)/\sqrt{2}; \tag{59}$$

$$|\Psi^+\rangle = (|01\rangle + |10\rangle)/\sqrt{2}, \quad |\Psi^-\rangle = (|01\rangle - |10\rangle)/\sqrt{2} \tag{60}$$

are entangled.

Although the notion of entanglement is mathematically simple, its physical interpretation is one of the main problems of modern quantum foundations. The common interpretation is that entanglement encodes quantum nonlocality, the possibility of action at the distance (between parts of a system in an entangled state). Such an interpretation implies the drastic change of all classical physical presentations about nature, at least about the microworld. In the probabilistic terms entanglement induces correlations which are too strong to be described by classical probability theory. (At least this is the common opinion of experts in quantum information theory and quantum foundations.) Such correlations violate the famous Bell inequality which can be derived only in classical probability framework. The latter is based on the use of a single probability space covering probabilistic data collected in a few incompatible measurement contexts.

Now consider a quantum state given by density operator ρ in H. This state is called separable if it can be factorized in the product of density operators in spaces H_k :

$$\rho = \otimes_{k=1}^{n} \rho_k, \tag{61}$$

otherwise the state ρ is called entangled. We remark that an interpretation of entanglement for mixed states is even more complicated than for pure states.

16 Violation of Formula of Total Probability in Two-Slit Experiment

Consider the famous two-slit experiment with the symmetric setting: the source of photons is located symmetrically with respect to two slits, Fig. 1.

Consider the following pair of observables a and b. We select a as the “slit passing observable,” i.e., $a = 0, 1$, see Fig. 1 (we use indexes 0, 1 to be close to qubit notation) and observable b as the position on the photo-sensitive plate, see Fig. 2. We remark that the b-observable has the continuous range of values, the position x on the photo-sensitive plate. We denote $\mathbf{P}(a = i)$ by $\mathbf{P}(i)$ $(i = 0, 1)$, and

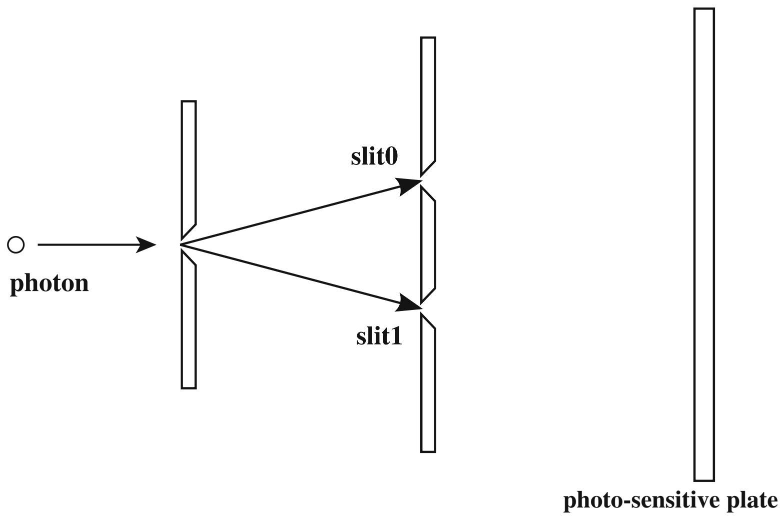


Fig. 1 Experimental setup

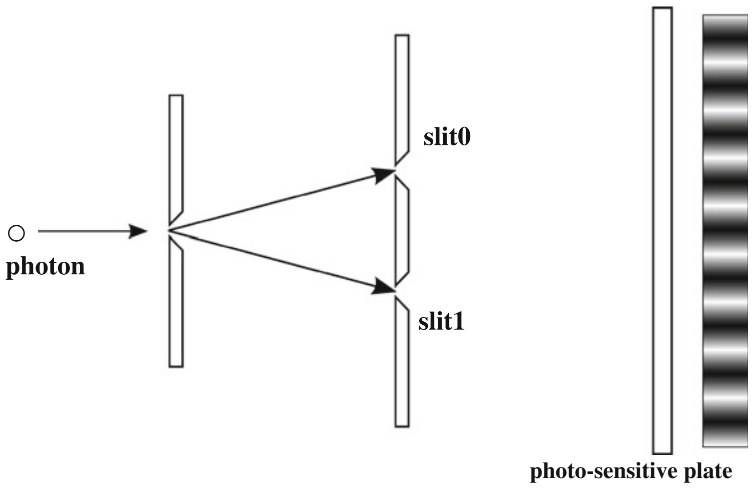


Fig. 2 Context with both slits is open

$\mathbf{P}(b = x)$ by $\mathbf{P}(x)$. Physically the a-observable corresponds to measurement of position (coarse grained to "which slit?") and the b-observable represents measurement of momentum.

In quantum foundational studies, various versions of the two-slit experiment have been successfully performed, not only with photons, but also with electrons and even with macroscopic molecules (by Zeilinger's group). All those experiment demonstrated matching with predictions of QM. Experimenters reproduce the interference patterns predicted by QM and calculated by using the wave functions.

The probability that a photon is detected at position x on the photo-sensitive plate is represented as

$$\mathbf{P}(x) = \left| \frac{1}{\sqrt{2}} \psi_0(x) + \frac{1}{\sqrt{2}} \psi_1(x) \right|^2 = \frac{1}{2} |\psi_0(x)|^2 + \frac{1}{2} |\psi_1(x)|^2 + |\psi_0(x)| \, |\psi_1(x)| \cos \theta, \quad (62)$$

where ψ_0 and ψ_1 are two wave functions, whose squared absolute values $|\psi_i(x)|^2$ give the distributions of photons passing through the slit $i = 0, 1$, see Figs. 3 and 4. Here we explored the rule of addition of complex probability amplitudes, a quantum analog of the rule of addition of probabilities. This rule is the direct consequence of the linear space structure of quantum state spaces.

The term

$$|\psi_0(x)| \, |\psi_1(x)| \cos \theta$$

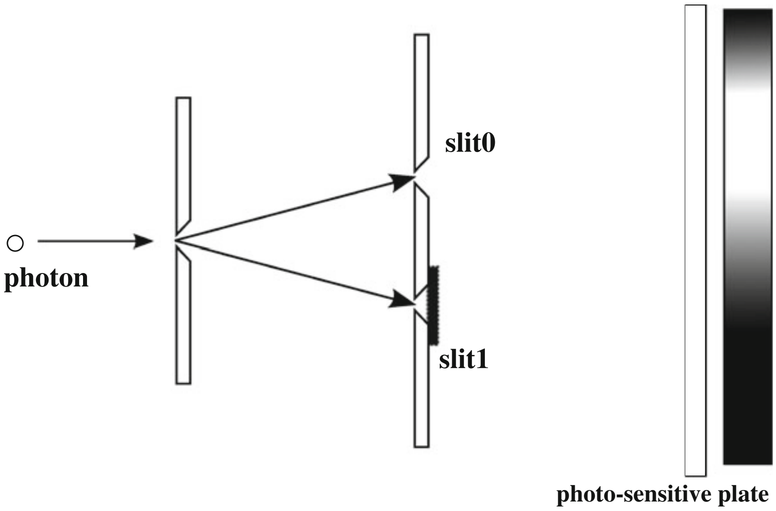


Fig. 3 Context with one slit is closed-I

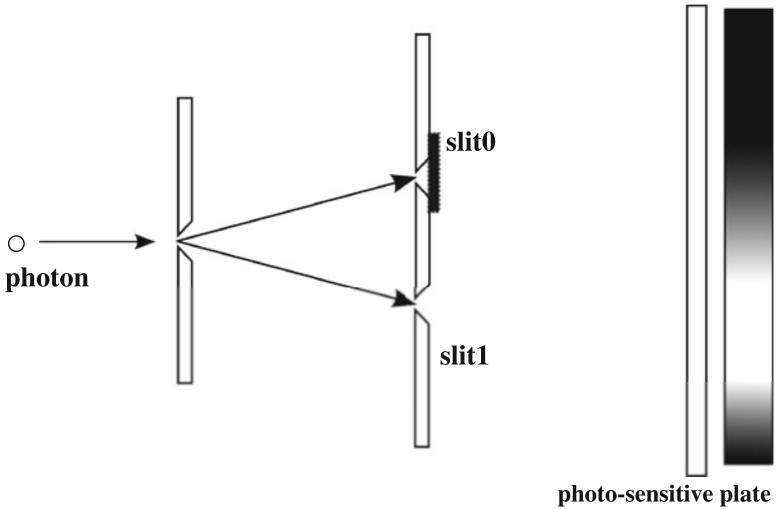


Fig. 4 Context with one slit is closed-II

implies the interference effect of two wave functions. Let us denote $|\psi_i(x)|^2$ by $\mathbf{P}(x|i)$, then Eq. (62) is represented as

$$\mathbf{P}(x) = \mathbf{P}(0)\mathbf{P}(x|0) + \mathbf{P}(1)\mathbf{P}(x|1) + 2\sqrt{\mathbf{P}(0)\mathbf{P}(x|0)\mathbf{P}(1)\mathbf{P}(x|1)}\cos\theta. \tag{63}$$

Here the values of probabilities $\mathbf{P}(0)$ and $\mathbf{P}(1)$ are equal to $1/2$ since we consider the symmetric settings. For general experimental settings, $\mathbf{P}(0)$ and $\mathbf{P}(1)$ can be taken as the arbitrary nonnegative values satisfying $\mathbf{P}(0) + \mathbf{P}(1) = 1$. In the above form, the classical probability law (FTP)

$$\mathbf{P}(x) = \mathbf{P}(0)\mathbf{P}(x|0) + \mathbf{P}(1)\mathbf{P}(x|1) \tag{64}$$

is violated, and the term of interference $2\sqrt{\mathbf{P}(x|0)\mathbf{P}(0)\mathbf{P}(x|1)\mathbf{P}(1)}\cos\theta$ specifies the violation.

The crucial point is that the two-slit experiment has the multi-contextual structure: $C_i, i = 0, 1$, only the ith slit is open, and C_{01}, both slits are open, see Figs. 3, 4, and 2. Comparison of possibilities is represented as comparison of the corresponding probability distributions $\mathbf{P}(x|i), \mathbf{P}(x)$. In the contextual notations they can be written as

$$p^b_{C_i}(x) \equiv \mathbf{P}(b = x|C_i),\ p^b_{C_{01}}(x) \equiv \mathbf{P}(b = x|C_{01}).$$

Here conditioning is not classical probabilistic event conditioning, but context conditioning: different contexts are mathematically represented by different Kolmogorov probability spaces. The general contextual probability theory including its representation in complex Hilbert space is presented in very detail in my monograph [37].

References

1. Aerts, D., Sozzo, S., & Tapia, J. (2012). A quantum model for the Ellsberg and Machina Paradoxes. *Quantum Interaction Lecture Notes in Computer Science, 7620*, 48–59.
2. Asano, M., Basieva, I., Khrennikov, A., Ohya, M., Tanaka, Y., & Yamato, I. (2015). Quantum information biology: From information interpretation of quantum mechanics to applications in molecular biology and cognitive psychology. *Foundations of Physics, 45*(10), 1362–1378.
3. Asano, M., Basieva, I., Khrennikov, A., & Yamato, I. (2017). A model of differentiation in quantum bioinformatics. *Progress in Biophysics and Molecular Biology, 130*(Part A), 88–98.
4. Asano, M., Basieva, I., Khrennikov, A., Ohya, M., & Tanaka, Y. (2017). A quantum-like model of selection behavior. *Journal of Mathematical Psychology, 78*, 2–12.
5. Asano, M., Khrennikov, A., Ohya, M., Tanaka, Y., & Yamato, I. (2014). Violation of contextual generalization of the Leggett-Garg inequality for recognition of ambiguous figures. *Physica Scripta, T163*, 014006.
6. Asano, M., Khrennikov, A., Ohya, M., Tanaka, Y., & Yamato, I. (2015). *Quantum adaptivity in biology: From genetics to cognition*. Heidelberg: Springer.
7. Ballentine, L. E. (1989). The statistical interpretation of quantum mechanics. *Reviews of Modern Physics, 42*, 358–381.
8. Ballentine, L. E. (1998). *Quantum mechanics: A modern development*. Singapore: WSP.
9. Ballentine, L. E. (2001). Interpretations of probability and quantum theory. In A. Y. Khrennikov (Ed.), *Foundations of probability and physics. Quantum probability and white noise analysis* (Vol. 13, pp. 71–84). Singapore: WSP.
10. Basieva, I., Khrennikova, P., Pothos, E. M., Asano, M., & Khrennikov, A. (2018). Quantum-like model of subjective expected utility. *Journal of Mathematical Economics, 78*, 150–162.
11. Birkhoff, J., & von Neumann, J. (1936). The logic of quantum mechanics. *Annals of Mathematics, 37*(4), 823–843.

12. Bohr, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physics Review, 48*, 696–702.
13. Bohr, N. (1938). The causality problem in atomic physics. In J. Faye & H. J. Folse (Eds.) (1987). *The philosophical writings of Niels Bohr*. Causality and Complementarity, Supplementary Papers (Vol. 4, pp. 94–121). Woodbridge: Ox Bow Press.
14. Bohr, N. (1987). *The philosophical writings of Niels Bohr* (Vol. 3). Woodbridge: Ox Bow Press.
15. Broekaert, J., Basieva, I., Blasiak, P., & Pothos, E. M. (2017). Quantum dynamics applied to cognition: A consideration of available options. *Philosophical Transactions of the Royal Society A, 375*(2016), 375–405.
16. Brukner, C., & Zeilinger, A. (1999). Malus' law and quantum information. *Acta Physica Slovaca, 49*(4), 647–652.
17. Brukner, C., & Zeilinger, A. (2009). Information invariance and quantum probabilities. *Foundations of Physics, 39*, 677–689.
18. Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge: Cambridge University Press.
19. Chiribella, G., D'Ariano, G. M. & Perinotti, P. (2010). Probabilistic theories with purification. *Physical Review A, 81*, 062348.
20. D' Ariano, G. M. (2007). Operational axioms for quantum mechanics. In: Adenier et al. (Eds.), *Foundations of probability and physics-3*. AIP conference proceedings (Vol. 889, pp. 79–105).
21. Dirac, P. A. M. (1995). *The principles of quantum mechanics*. Oxford: Clarendon Press.
22. Feynman, R., & Hibbs, A. (1965). *Quantum mechanics and path integrals*. New York: McGraw-Hill.
23. Feynman, R. P. (1951). The concept of probability in quantum mechanics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability* (pp. 533–541). California: University of California Press.
24. Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., & van Rijsbergen, K. (2010). Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *IIiX'10* (pp. 115–124).
25. Fuchs, C. A. (2002). Quantum mechanics as quantum information (and only a little more). In A. Khrennikov (Ed.), *Quantum theory: Reconsideration of foundations. Series mathematical modeling* (Vol. 2, pp. 463–543). Växjö: Växjö University Press.
26. Fuchs, C. A., Mermin, N. D., & Schack, R. (2014). An Introduction to QBism with an application to the locality of quantum mechanics. *American Journal of Physics, 82*, 749.
27. Haven, E., & Khrennikov, A. (2009). Quantum mechanics and violation of the sure-thing principle: The use of probability interference and other concepts. *Journal of Mathematical Psychology, 53*, 378–388.
28. Haven, E., & Khrennikov, A. (2013). *Quantum social science*. Cambridge: Cambridge University Press.
29. Khrennikov, A. (2009). *Interpretations of probability* (2nd ed.). Tokyo/Berlin: VSP International SC Publisher/De Gruyter.
30. Khrennikov, A. Y. (2001). Linear representations of probabilistic transformations induced by context transitions. *Journal of Physics A: Mathematical and General, 34*, 9965–9981.
31. Khrennikov, A. Y. (2001). Origin of quantum probabilities. In A. Khrennikov (Ed.), *Foundations of probability and physics* (pp. 180–200). Singapore: Växjö-2000/WSP.
32. Khrennikov, A. (2004). *Information dynamics in cognitive, psychological, social, and anomalous phenomena. Fundamental theories of physics*. Dordrecht: Kluwer.
33. Khrennikov, A. (2004). Växjö interpretation-2003: Realism of contexts. In *Proceeding International Conference Quantum Theory: Reconsideration of Foundations. Mathematical Modelling in Physics and Engineering, and Cognitive Science* (Vol. 10, pp. 323–338). Växjö: Växjö University Press.
34. Khrennikov, A. (2005). Reconstruction of quantum theory on the basis of the formula of total probability. In *AIP Conference Proceeding of the Foundations of Probability and Physics—3* (Vol. 750, pp. 187–218). Melville: American Institute of Physics.

35. Khrennikov, A. (2005). The principle of supplementarity: A contextual probabilistic viewpoint to complementarity, the interference of probabilities, and the incompatibility of variables in quantum mechanics. *Foundations of Physics, 35*(10), 1655–1693.
36. Khrennikov, A. (2008). Algorithm for quantum-like representation: Transformation of probabilistic data into vectors on Bloch's sphere. *Open Systems and Information Dynamics, 15*, 223–230.
37. Khrennikov, A. (2009). *Contextual approach to quantum formalism*. Berlin: Springer.
38. Khrennikov, A. (2010). *Ubiquitous quantum structure: From psychology to finances*. Berlin: Springer.
39. Khrennikov, A. (2015). Quantum-like model of unconscious-conscious dynamics. *Frontiers in Psychology, 6*. Art. 997.
40. Khrennikov, A. (2015). Towards information lasers. *Entropy, 17*(10), 6969–6994.
41. Khrennikov, A. (2015). Quantum version of Aumann's approach to common knowledge: Sufficient conditions of impossibility to agree on disagree. *Journal of Mathematical Economics, 60*, 89–104.
42. Khrennikov, A. (2016). Analog of formula of total probability for quantum observables represented by positive operator valued measures. *International Journal of Theoretical Physics, 55*, 859–3874.
43. Khrennikov A. Yu. (2016). *Probability and randomness: Quantum versus classical*. Singapore: World Scientific.
44. Khrennikov, A. (2016). Quantum Bayesianism as the basis of general theory of decision-making. *Philosophical Transactions of the Royal Society A, 374*, 20150245.
45. Khrennikova, P. (2016, April). Quantum dynamical modeling of competition and cooperation between political parties: The coalition and non-coalition equilibrium model. *Journal of Mathematical Psychology, 71*, 39–50.
46. Khrennikov, A., & Basieva, I. (2014). Quantum model for psychological measurements: From the projection postulate to interference of mental observables represented as positive operator valued measures. *NeuroQuantology, 12*, 324–336.
47. Khrennikova, P., & Haven, E. (2016). Instability of political preferences and the role of mass-media: a dynamical representation in a quantum framework. *Philosophical Transactions of the Royal Society A, 374*, 20150106.
48. Khrennikova, P., Haven, E., & Khrennikov, A. (2014) An application of the theory of open quantum systems to model the dynamics of party governance in the US political System. *International Journal of Theoretical Physics, 53*(4), 1346–1360.
49. Kolmolgoroff, A. N. (1933). *Grundbegriffe der wahrscheinlichkeitsrechnung*. Berlin: Springer.
50. Kolmolgorov, A. N. (1956). *Foundations of the probability theory*. New York: Chelsea Publishing Company.
51. Melucci, M. (2013). Deriving a quantum information retrieval basis. *The Computer Journal, 56*(11), 1279–1291.
52. Melucci, M. (2015). *Introduction to information retrieval and quantum mechanics*. Berlin: Springer.
53. Plotnitsky, A. (2012). *Niels Bohr and complementarity: An introduction*. Berlin: Springer.
54. Plotnitsky, A., & Khrennikov, A. (2015). Reality without realism: On the ontological and epistemological architecture of quantum mechanics. *Foundations of Physics, 45*(10), 269–1300.
55. Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violation of rational decision theory. *Proceedings of the Royal Society B, 276*, 2171–2178.
56. Song, D., Lalmas, M., van Rijsbergen, C. J., Frommholz, I., Piwowarski, B., Wang, J., et al. (2010). How quantum theory is developing the field of information retrieval? In *Quantum informatics for cognitive, social, and semantic processes: Papers from the AAAI fall symposium (FS-10-08)* (pp. 105–108).
57. Sozzo, S. (2014). A Quantum probability explanation in Fock space for borderline contradictions. *Journal of Mathematical Psychology, 58*(1), 1–12.

58. van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge: Cambridge University Press.
59. van Rijsbergen, C. J. (2011). What is Quantum information retrieval? In G. Amati, F. Crestani, (Eds.), *Advances in information retrieval theory. ICTIR 2011. Lecture notes in computer science* (Vol. 6931). Berlin: Springer.
60. von Neuman, J. (1955). *Mathematical foundations of quantum mechanics*. Princeton: Princeton University Press.
61. von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Representing Words in Vector Space and Beyond

Benyou Wang, Emanuele Di Buccio, and Massimo Melucci

Abstract Representing words, the basic units in language, is one of the most fundamental concerns in Information Retrieval, Natural Language Processing (NLP), and related fields. In this paper, we reviewed most of the approaches of word representation in vector space (especially state-of-the-art word embedding) and their related downstream applications. The limitations, trends and their connection to traditional vector space based approaches are also discussed.

Keywords Word representation · Word embedding · Vector space

1 Introduction

This volume illustrates how quantum-like models can be exploited in Information Retrieval (IR) and other decision making processes. IR is a special and important instance of decision making because, when searching for information, the users of a retrieval system express their information needs through behavior (e.g., click-through activity) or queries (e.g., natural language phrases), whereas a computer system decides about the relevance of documents to the user's information need. By nature, IR is inherently an interactive activity which is performed by a user accessing the collections managed by a system through very interactive devices. These devices are immersed in a highly dynamic context where not only does the user's queries rapidly evolve but the collections of documents such as news or magazine articles also use words with different meanings. The main link between the "quantumness" of these models and IR is established by the vector spaces, which have for a long time been utilized to design modern computerized systems such as the search engines and they are currently the foundation of the most advanced methods for searching for multimedia information.

B. Wang (✉) · E. Di Buccio · M. Melucci
Department of Information Engineering, University of Padova, Padova, Italy
e-mail: wang@dei.unipd.it; dibuccio@dei.unipd.it; massimo.melucci@unipd.it

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_5

Whatever the mathematical model or the retrieval function, documents and queries are mathematically represented as elements of sets, while the sets are labeled by words or other document properties. Queries, which are the most used data for expressing information needs, are sets or sequences of words or they are sentences expressed in a natural language; queries are oftentimes very short (e.g., one word) or occasionally much longer (e.g., a text paragraph). It is a matter of fact that the Boolean models for IR by definition view words as document sets and answer search queries with document sets obtained by set operators; moreover, the probabilistic models are all inspired to the Kolmogorov theory of probability, which is related to Boole's theory of sets; in addition, the traditional retrieval models based on vector spaces are eventually a means to provide a ranking or a measure to sets because they assign a weight to words and then to documents in the sets labeled by the occurring words. The implementation of content representation in terms of keywords and posting lists reflects the view of words as sets of documents and the view of retrieval operations as set operators. In this chapter, we will explain that a document collection can be searched by vectors embedding different words together, instead of by distinct words, by using the ultimate logic of *vector spaces*, instead of sets.

Representing words is fundamental for tasks which involve sentences and documents. Word embedding is a family of techniques that has recently gained a great deal of attention and aims at learning vector representation of words that can be used in these tasks. Generally speaking, embedding mainly consists in adopting a mapping, in which a fixed-length vector is typically used to encode and represent an entity, e.g., word, document, or a graph. Technically, in order to embed an object X in another object Y, the embedding is an injective and structure-preserving map $f : X \rightarrow Y$, e.g., user/item embedding [6] in item recommendation, network embedding [23], feature embedding in manifold learning [89], and word embedding. In this chapter, we will focus on word embedding techniques, which embed words in a low-dimensional vector space.

Word embedding is driven by the *Distributional Hypothesis* [33, 38], which assumes that linguistic items which occur in similar contexts should have similar meanings. Methods for modeling the distributional hypothesis can be mainly divided into the following categories:

- Vector-space models in Information Retrieval, e.g., [121], or representation in Semantic Spaces [67]
- Cluster-based distributional representation [17, 63, 79]
- Dimensionality reduction (matrix factorization) for document-word/word-word/word-context co-occurring matrix, also known as Latent Semantic Analysis (LSA) [24]
- Prediction based word embedding, e.g., using neural network-based approaches.

LSA was proposed to extract descriptors that capture word and document relationships within one single model [24]. In practice, LSA is an application of Singular Value Decomposition (SVD) to a document-term matrix. Following LSA, Latent Dirichlet Allocation (LDA) aims at automatically discovering the main topics in a document corpus. A corpus is usually modeled as a probability distribution over a shared set of topics; these topics in turn are probability distributions over words, and each word in a document is generated by the topics [12]. This paper focuses on the geometry provided by vector spaces, yet is also linked to topic models, since a probability distribution over documents or features is defined in a vector space, the latter being a core concept of the quantum mechanical framework applied to IR [68, 69, 110].

With the development of computing ability for exploiting large labeled data, neural network-based word embedding tends to be more and more dominant, e.g., Computer Vision (CV) and Natural Language Processing. In the NLP field, neural network-based word embedding was firstly investigated by Bengio et al. [7] and further developed by [21, 75]. Word2vec [70][1] adopts a more efficient way to train word embedding, by removing non-linear layers and other tricks, e.g., hierarchical softmax and negative sampling. In [70] the authors also discussed the *additive compositional structure*, which denotes that word meanings can be composited with the addition of their corresponding vectors. For example, $\boldsymbol{king} - \boldsymbol{man} = \boldsymbol{queen} - \boldsymbol{women} = \boldsymbol{royal}$. This capability of capturing relationships among words was further discussed in [35] where a theoretical justification was provided. More importantly, Mikolov et al. [70] published open-source well-trained general word vectors, which made word embedding easy to use in various tasks.

In order to intuitively show the word vectors, some selected words (52 words about animals and 110 words about colors) are visualized in a 2-dimensional plane (as shown in Fig. 1) from one of the most popular Glove word vectors,[2] in which the position of the word is according to the reduced vector through a dimension reduction approach called T-SNE. It is shown that all the words are nearly clustered into two groups about colors and animals, respectively. For example, the word vectors of "rat" and "dog" are close to the word "cat," which is intuitively consistent to the Distributional Hypothesis since they ("cat" and "rat," or "cat" and "dog") may co-occur together with high frequencies.

Word embedding provides a more flexible and fine-grained way to capture the semantics of words, as well as to model the semantic composition of bigger-granularity units, e.g., from words to sentences or documents [71]. Some applications of word embedding will be discussed in Sect. 3. Although word embedding techniques and related neural network approaches have been successfully used in different IR and NLP tasks, they have some limitations, e.g., the polysemy and out-of-vocabulary problems. These issues have motivated further research in word embedding; Sect. 4.2 will discuss some of the current trends in word embedding that aim at addressing these issues. Moreover, we will discuss the link between the word vector representations and state-of-the-art approaches in modeling thematic structures.

[1] https://code.google.com/archive/p/word2vec/.

[2] The words vectors are downloaded from http://nlp.stanford.edu/data/glove.6B.zip, with 6B tokens, 400K uncased words, and 50-dimensional vectors.

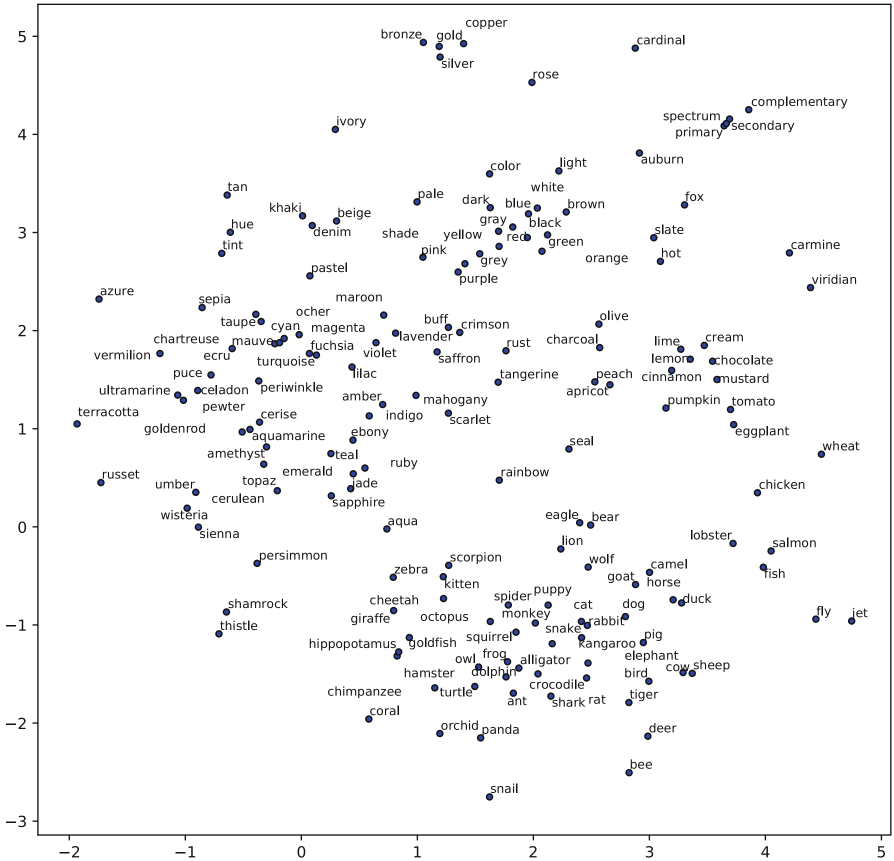


Fig. 1 The visualization of some selected words

2 Background

2.1 Distributional Hypothesis

Word embedding is driven by the *Distributional Hypothesis* [38]. The core of distributional hypothesis states that linguistic items with similar distributions have similar meanings and hence words with similar distributions should have similar representations. The distributional property is usually induced from document or textual neighborhoods (like sliding windows).

Some of the methods relying on the Distributional Hypothesis and the basic idea underlying them are reported below:

– Language model $p(w_k|w_{k-t}, w_{k-t+1}, \ldots w_{k-1})$: predicts the current word using previous words [7].
– Sequential scoring $p(w_{k-t}, w_{k-t+1}, \ldots w_k)$: predicts whether the given sentence is a legal one [21].
– Skip-gram $p(w_k| \, \forall w_i \in \{w_i| \, abs(k-i) < t\} \,)$: predicts a co-occurring word for each word [70].
– CBOW $p(w_k|w_{k-t}, w_{k-t+1}, \ldots w_{k-1}, \ldots w_{k+t})$: predicts a target word with context words (both previous ones and following ones) [70].
– Glove $p(\#(w_i, w_j)_{window}|w_i, w_j)$: predicts the co-occurring count between a word pair [78].

2.2 A Brief History of Word Embedding

While the *Distributional Hypothesis* was proposed many decades ago, the techniques of word embedding trained in a neural network has a much shorter history of about one and half decades [7], as mentioned in Sect. 1. Some typical ways to generate word vectors are introduced below.

NNLM The Neural Network Language Model (NNLM) [7] preliminarily aims to build a language model, while learning word embedding is not the main target. However, this is the first work in learning word vectors in a neural network (Fig. 2).

C&W The Collobert and Weston (C&W) approach was proposed in [21] in order to predict the fluency of a given sequence—see Fig. 3. One of the tasks in [21] assigns language modeling as a simple binary classification task: “if the word in the middle of the input window is related to its context or not” [21].

Skip-Gram Skip-gram balances a trade-off between performance and simplicity. As shown in Fig. 4, Skip-gram uses a word to predict one of its neighboring words.

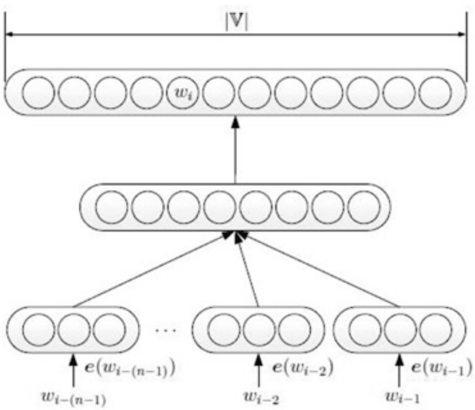


Fig. 2 NNLM concatenates all the word vectors in a sentence and then predicts the next word. → refers to the information flow in the forward neural network, while the circle denotes the neurons in the network. $|V|$ is the size of the word vocabulary [58]

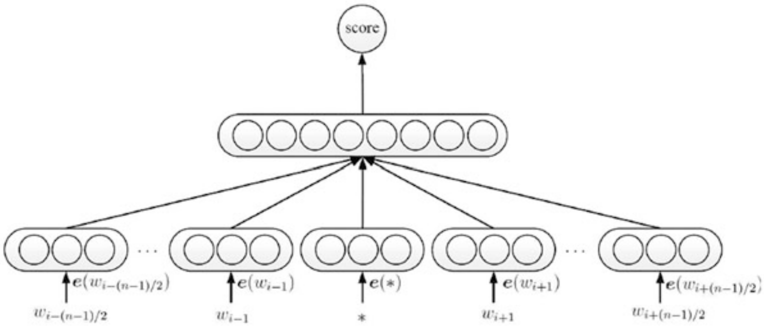


Fig. 3 C&W concatenates all the word vectors to predict whether it is a natural sentence or if it has replaced the center word with a random word [58]

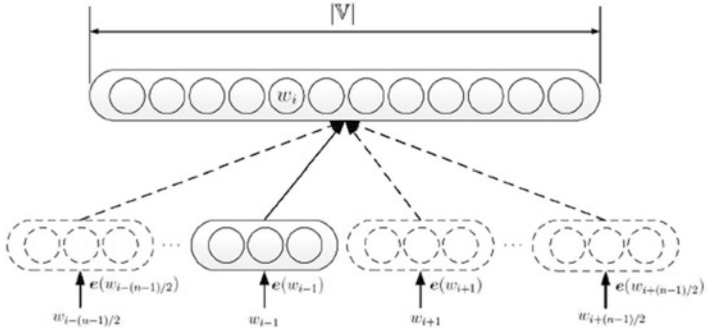


Fig. 4 Skip-gram directly uses one word to predict its neighboring word [58]

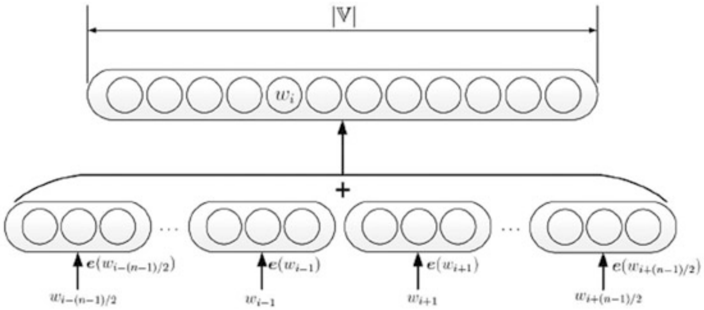


Fig. 5 CBOW uses the average embedding of the contextual words to predict the target word, where the contextual words are surrounded by the target word [58]

CBOW As shown in Fig. 5, CBOW uses context words to predict the current word. The difference between Skip-gram and CBOW is that in order to predict the target word, CBOW uses many words as the context, while Skip-gram uses only one neighboring word.

Glove Another popular word embedding named Glove[3] [78] takes advantage of global matrix factorization and local context window methods. It is worth mentioning that [60] explains that the Skip-gram with negative sampling derives the same optimal solution as matrix (Point-wise Mutual Information (PMI)) factorization.

3 Applications of Word Embedding

According to the input and output objects, we will discuss word-level applications in Sect. 3.1, sentence-level applications in Sect. 3.2, pair-level applications in Sect. 3.3, and seq2seq generation applications in Sect. 3.4. These applications can be the benchmarks to evaluate the quality of word embedding, as introduced in Sect. 3.5.

3.1 Word-Level Applications

Based on the learned word vector from a large-scale corpus, the word-level property can be inferred. Regarding *single-word level property*, word sentiment polarity is one of the typical properties. Word-pair properties are more common tasks, like word similarity and word analogy.

The advantage of word embedding is that: all the words, even from a complicated hierarchical structure like WordNet [31],[4] are embedded in a single word vector, thus leading to a very simple data structure and easy incorporation with a downstream neural network. Meanwhile, this simple data structure, namely a word-vector mapping, also provides some potential to share different knowledge from various domains.

3.2 Sentence-Level Application

Regarding sentence-level applications, the two typical tasks are sentence classification and sequential labeling, depending on how many labels the task needs. For a given sentence, there is only one final label for the whole sentence for text classification, where the number of labels in the sequential labeling is related to the number of tokens in the sentence (Fig. 6).

[3] https://nlp.stanford.edu/projects/glove/.

[4] An example of hierarchical structures is shown at the following address: http://people.csail.mit.edu/torralba/research/LabelMe/wordnet/test.html.

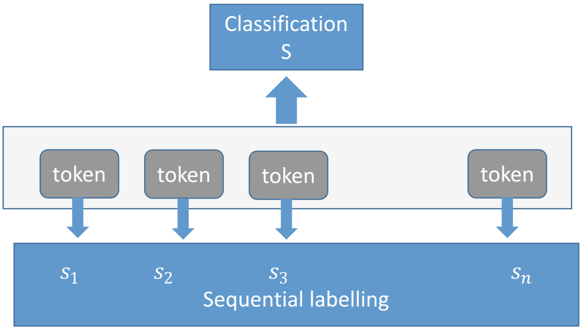


Fig. 6 Sentence-level applications: sentence classification and sequential labeling

Sentence Classification Sentence classification aims to predict the possible label for a given sentence, where the label can be related to the topic, the sentimental polarity, or whether the mail is spam. Text classifications were previously overviewed by Zhai [1], who mainly discussed the traditional textual representation. To some extent, trained word embedding from a large-scale *external* corpus (like Wikipedia pages or online news) is commonly used in IR and NLP tasks like text classification. Especially for a task with limited labeled data, in which it is impossible to train effective word vectors (usually with one hundred thousand parameters that need to be trained) due to the limited corpus, pre-trained embedding from a large-scale external corpus could provide general features. For example, average embedding (or with a weighted scheme) could be a baseline for many sentence representations and even document representations. However, due to the original error for the embedding training process in the external corpus and the possible domain difference between the current dataset and external corpus, adopting the embedding as features usually will not achieve significant improvement over traditional bag-of-word models, e.g., BM25 [88].

In order to solve this problem, the word vectors trained from a large-scale external corpus are only adopted as the initial value for the downstream task [51]. Generally speaking, all the parameters of the neural network are trained from scratch with a random or regularized initialization. However, the scale of the parameter in the neural network is large and the training samples may be small. Moreover, the trained knowledge from another corpus is expected to be used in a new task, which is commonly used in Computer Vision (CV) [41]. In an extreme situation, the current dataset is large enough to implicitly train the word embedding from scratch; thus, the effect of pre-initial embedding could be of little importance.

Firstly, multi-layer perception is adopted over the embedding layers. Kim et al. [51] first proposed a CNN-based neural network for sentence classification as shown in Fig. 7. The other typical neural networks named Recurrent Neural Network (and its variant called Long and Short Term Memory (LSTM) network [43] as shown in Fig. 8) and Recursive Neural Network [36, 81], which naturally process sequential sentences and tree-based sentences, are becoming more and more popular. In particular, word embedding with LSTM encoder–decoder architecture

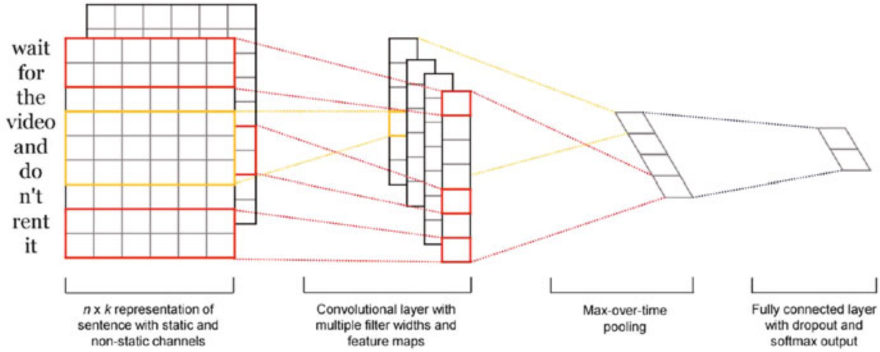


Fig. 7 CNN for sentence modeling [52] with convolution structures and max pooling

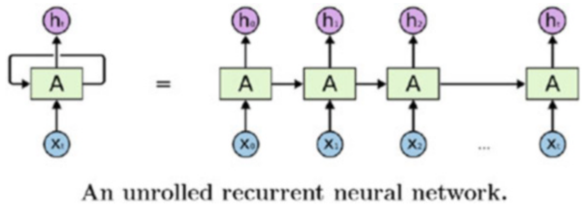


Fig. 8 LSTM. The left subfigure shows a recurrent structure, while the right one is unfolded over time

[3, 18] outperformed the classical statistic machine translation,[5] which dominates machine translation approaches. Currently, the industrial community like Google adopts completely neural machine translation and abandons statistical machine translation.[6]

Sequential Labeling Sequence labeling aims to classify each item of a sequence of observed value, with the consideration of the whole context. For example, Part-Of-Speech (POS) tagging, also called word-category disambiguation, is the process of assignment of each word in a text (corpus) to a particular part-of-speech label (e.g., noun and verb) based on its context, i.e., its relationship with adjacent and related words in a phrase or sentence. Similar to the POS tagging, the segment tasks like Named Entity Recognition (NER) and word segment can also be implemented in a general sequential labeling task, with definitions of some labels like begin label (usually named "B"), intermediate label (usually named "O"), and end label (usually named "E"). The typical architecture for sequence labeling is called BiLSTM-CRF [46, 59], which is based on bidirectional LSTMs and conditional random fields, as shown in Fig. 9.

[5] http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf.

[6] https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

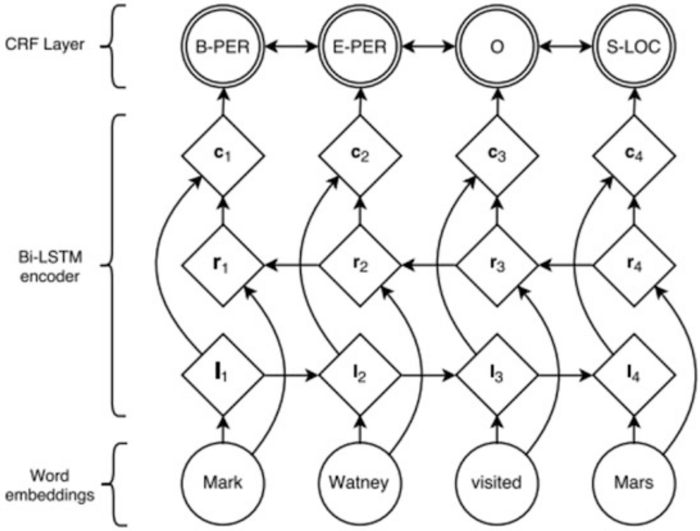


Fig. 9 LSTM-CRF for named entity recognition [59]

Document-Level Representation Similar to the methods for sentence-level representation, a document with mostly multiple sentences, which can also be considered a long "sentence," needs an adaption for more tokens. A document mostly consists in multiple sentences. If we interpret a document as a long sentence, we can use the same approaches proposed for the sentence-level applications while taking into account the fact that there are more tokens. For example, a hierarchical architecture is usually adopted for document representation, especially in RNN, as shown in Fig. 10. Generally speaking, all the sentence-level approaches can be used in document-level representation, especially if the document is not so long.

3.3 Sentence-Pair Level Application

The difference between sentence applications and sentence-pair applications is the extra *interaction* module (we call it a matching module), as shown in Fig. 11. Evaluating the relationship between two sentences (or a sentence pair) is typically considered a matching task, e.g., information retrieval [73, 74, 129], natural language inference [14], paraphrase identification [27], and question answering. It is worth mentioning that the Reading Comprehension (RC) task can also be a matching task (especially question answering) when using an extra context, i.e., a passage for background knowledge, while the question answering (answer selection) does not have specific context. In the next subsection, we will introduce the Question Answering task and Reading Comprehension task.

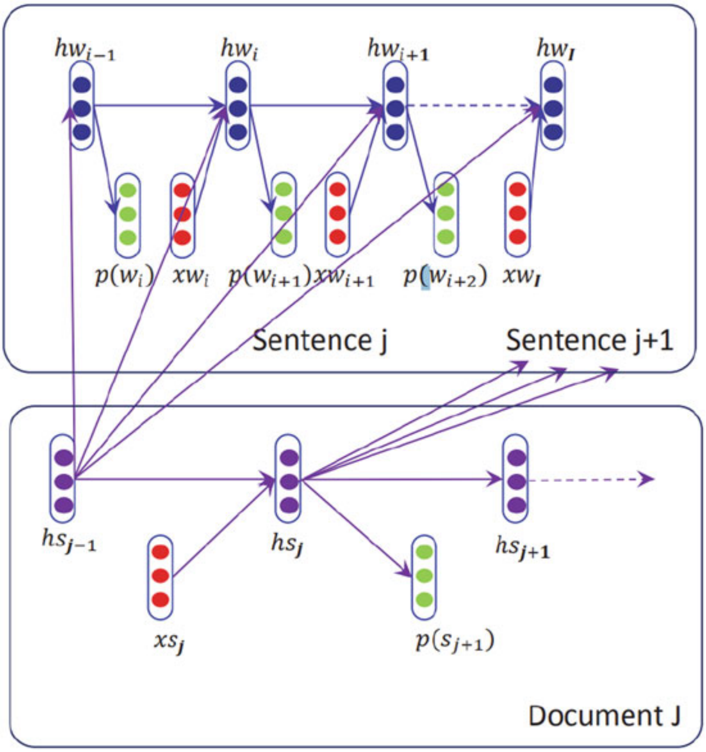


Fig. 10 Hierarchical recurrent neural network [64]

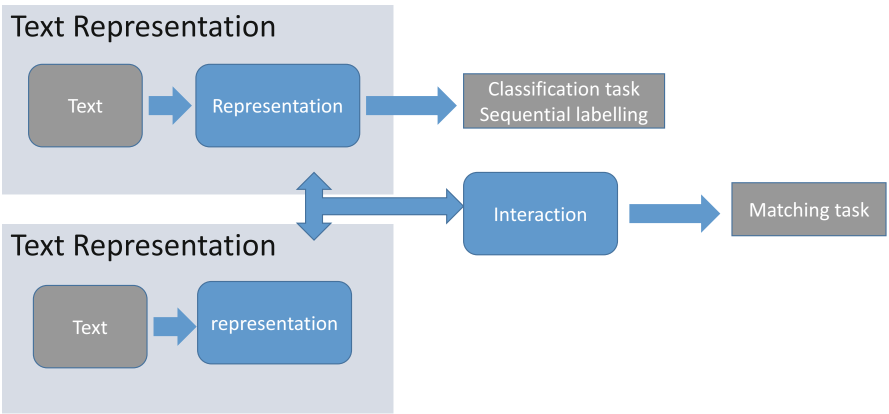


Fig. 11 The figure shows that the main difference between a sentence-pair task and a sentence-based task is that there is one extra interaction for the matching task

Fig. 12 A demo of SQuAD dataset [85]

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Question Answering Differently from *expert systems* with structured knowledge, question answering in IR is more about retrieval and ranking tasks in limited unstructured document candidates. In some literature, reading comprehension is also considered a question answering task like SQuAD QA. Generally speaking, reading comprehension is a question answering task in a specific context like a long document with some internal phrases or sentences as answers, as shown in Fig. 12. Table 1 reports current popular QA datasets.

In order to compare the neural matching model and non-neural models, we focus on TREC (answer selection), which has limited answer candidates, instead of an unstructured document as context in reading comprehension. Some matching methods are shown in Table 2, which mainly refers to the ACL wiki page.[7]

3.4 *Seq2seq Application*

Seq2seq is a kind of task with both input and output as sequential objects, like a machine translation task. It mainly uses an encoder–decoder architecture [19, 100] and further attention mechanisms [3], as shown in Fig. 13. Both the encoder and decoder can be implemented as RNN [19], CNN [34], or only attention mechanisms (i.e., Transformer [111]).

[7] https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art).

Table 1 Popular QA dataset

Dataset	Characteristics	Main institution	Venue
TREC QA [119][a]	Open-domain question answering	CMU	EMNLP 2007
Insurance QA [32]	Question answering for insurance	IBM Watson	ASRU 2015
Wiki QA [123]	Open-domain question answering	MS	EMNLP 2015
Narrative QA [53]	Reading Comprehension	DeepMind	TACL 2018
SQuAD 1.0 [85]	Questions for machine comprehension	Standford	EMNLP 2016
MS Marco [76]	Human-generated machine reading	MS.	NIPS 2016
NewsQA [107, 108]	Reading comprehension	Maluuba	RepL4NLP 2017
TriviaQA [48]	Reading comprehension distantly supervised labels	Allen AI	ACL 2017
SQA [47]	Sequential question answering	U. of Maryland & MS.	ACL 2017
CQA [102]	QA with knowledge base of web	Tel-Aviv university	NAACL 2018
CSQA [92]	Complex sequential QA	IBM	AAAI 2018
QUAC [20][b]	Question answering in context	Allen AI	EMNLP 2018
SQuAD 2.0 [84]	SQuAD with unanswered questions	Standford	ACL 2018
CoQA [87][c]	Conversational question answering	Standford	Aug. 2018
Natural questions [57]	Natural questions in Google search	Google	TACL 2019

The frequent publishing of QA datasets demonstrates that the academic community is paying more and more attention to this task. Almost all the researchers in this community tend to use word embedding-based neural networks for this task

[a] http://cs.stanford.edu/people/mengqiu/data/qg-emnlp07-data.tgz

[b] http://quac.ai/

[c] https://stanfordnlp.github.io/coqa/

3.5 Evaluation

The basic evaluations of word embedding techniques are based on the above applications [94], e.g., word-level evaluation and downstream NLP tasks like those mentioned in the last section, as shown in [58]. Especially for a downstream task, there are two common ways to use word embedding, namely as fixed features or by treating it only as initial weights and fine-tuning it. We mainly divide it into two part of evaluations, i.e., context-free word properties and embedding-based downstream NLP tasks, while the latter may involve the context and the embedding can be fine-tuned.

Word Property Examples of the context-free word properties include word polarity classification, word similarity, word analogy, and recognition of synonyms and antonyms. In particular, one of the typical tasks is called an analogy task [70], which

Table 2 State-of-the-art methods for sentence selection, where the evaluation relies on the TREC QA dataset

Algorithm	Reference	MAP	MRR
Mapping dependencies trees [82]	AI and math Symposium 2004	0.419	0.494
Dependency relation [22]	SIGIR 2005	0.427	0.526
Quasi-synchronous grammar [119]	EMNLP 2007	0.603	0.685
Tree edit models [42]	NAACL 2010	0.609	0.692
Probabilistic tree edit models [118]	COLING 2010	0.595	0.695
Tree edit distance [124]	NAACL 2013	0.631	0.748
Question classifier, NER, and tree kernels [95]	EMNLP 2013	0.678	0.736
Enhanced lexical semantic models [126]	ACL 2013	0.709	0.770
DL with bigram+count [128]	NIPS 2014 DL workshop	0.711	0.785
LSTM—three-layer BLSTM+BM25 [116]	ACL 2015	0.713	0.791
Architecture-II [32, 45]	NIPS 2014	0.711	0.800
L2R + CNN + overlap [96]	SIGIR 2015	0.746	0.808
aNMM: [122] attention-based neural matching model	CIKM 2016	0.750	0.811
Holographic dual LSTM architecture [104]	SIGIR 2017	0.750	0.815
Pairwise word interaction modeling [40]	NAACL 2016	0.758	0.822
Multi-perspective CNN [39]	EMNLP 2015	0.762	0.830
HyperQA (hyperbolic embeddings) [103]	WSDM 2018	0.770	0.825
PairwiseRank + multi-perspective CNN [86]	CIKM 2016	0.780	0.834
BiMPM [120]	IJCAI 2017	0.802	0.875
Compare-aggregate [8]	CIKM 2017	0.821	0.899
IWAN [97]	EMNLP 2017	0.822	0.889
IWAN + sCARNN [106]	NAACL 2018	0.829	0.875
NNQLM [131]	AAAI 2018	0.759	0.825
Multi-cast attention networks (MCAN) [105]	KDD 2018	0.838	0.904

Recent papers about TREC QA used embedding-based neural network approaches, while previous ones were based on some traditional methods like IR approaches and edit distance

mainly targets both the syntactic and semantic analogies. For instance, “man is to woman” is semantically similar to “king is to queen,” while “predict is to predicting” is syntactically similar to “dance is to dancing.” Word Embedding methods achieve good performance in the above word-level tasks, which demonstrates that the word embedding can capture the basic semantic and syntactic properties of the word.

Downstream Task If word embedding is used in a context, which means we consider each word in a phrase or sentence for a specific target, we can train the word embedding by using the labels of the specific task, e.g., sequential labeling, text classification, text matching, and machine translation. These tasks are divided by the pattern of input and output, shown in Table 3.

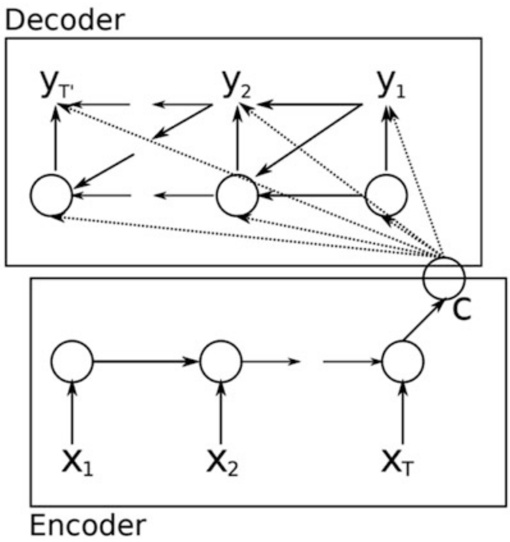


Fig. 13 An illustration of the proposed Seq2seq (RNN Encoder–Decoder)

Table 3 The difference of the downstream tasks

Algorithm	Input	Output	Typical tasks	Typical models
Text classification	S	$\mathcal{R}$	Sentiment analysis, topic classification	Fastext/CNN/RNN
Text matching	(S_1, S_2)	$\mathcal{R}$	QA, reading comprehension	aNMM,DSSM
Sequential labeling	S	$\mathcal{R}^{\|S\|}$	POS, word segmentation, NER	LSTM-CRF
Seq2Seq	S_1	S_2	machine translation, abstraction	LSTM/Transformer encoder–decoder

Generally speaking, the tasks for the word properties can partially reflect the quality of the word embedding. However, the final performance in the downstream tasks may vary. It is more reasonable to directly assess it in the real-world downstream tasks as shown in Table 3.

4 Reconsidering Word Embedding

Some limitations and trends of word embedding are introduced in Sects. 4.1 and 4.2. We also try to discuss the connections between word embedding and topic models in Sect. 4.3. In Sect. 4.4, the dynamic properties of word embedding are discussed in detail.

4.1 Limitations

Limitation of Distributional Hypothesis The first concern directly targets the effectiveness of the distributional hypothesis. Lucy and Gauthier [66] find that while word embeddings capture certain *conceptual* features such as "is edible" and "is a tool," they do not tend to capture *perceptual features* such as "is chewy" and "is curved," potentially because the latter are not easily inferred from distributional semantics alone.[8]

Lack of Theoretical Explanation Generally, humans perceive the words with various aspects other than only the semantic aspect, e.g., sentimental polarity and semantic hierarchy like WordNet. Thus, mapping a word to a real-valued vector is a practical but preliminary method, which leads to limited hints for humans to understand. For a given word vector, it is hard for humans to know what exactly the word means; the scalar value of each element in a word vector does not provide too much physical meaning. Consequently, it is difficult to interpret obtained vector space from the human point of view.

Polysemy Problem Another problem with word embeddings is that they do not account for *polysemy*, instead assigning exactly one vector per surface form. Several methods have been proposed to address this issue. For example, Athiwaratkun and Wilson [2] represent words not by single vectors, but by Gaussian probability distributions with multiple modes—thus capturing both uncertainty and polysemy. Upadhyay et al. [109] leverage multi-lingual parallel data to learn multi-sense word embeddings, for example, the English word "bank," which can be translated into both the French words *banc* and *banque* (evidence that "bank" is polysemous), and help distinguish its two meanings.

Out-Of-Vocabulary Problem With a pre-trained word embedding, some words may not be found in the vocabulary of the pre-trained word vectors, that is, the Out-Of-Vocabulary (OOV) problem. If there are many OOV words, the final performance decreases largely due to the fact that we use a partial initialization from the given word vectors, while other words are randomly initialized, instead of initializing all the weights. This happened more frequently in some professional domains, like medicine text analysis, since it is not easy to find some professional words in a general corpus like Wikipedia.

Semantic Change Over Time One of the limitations of most word embedding approaches is that they assume that the meaning of a word does not change over time. This assumption can be a limitation when considering corpora of historic texts or streams of text in newspapers or social media. Section 4.4 will discuss some recent works which aim to explicitly include the temporal dimensions in order to capture how the word meaning changes over time.

[8] http://www.abigailsee.com/2017/08/30/four-deep-learning-trends-from-acl-2017-part-1.html.

4.2 Trends

Interpretability One of the definitions of "interpretability" is proposed by Lipton [65]. In particular, Lipton [65] identifies two broad approaches to interpretability: *post-hoc explanations* and *transparency*. Post-hoc explanations take a learned model and draw some useful insights from it; typically these insights provide only a partial or **indirect** explanation of how the model works. The typical examples are visualization (e.g., in machine translation [26]) and transfer learning.

Transparency asks more **directly** "how does the model work?" and seeks to provide some way to understand the core mechanisms of the model itself. As Manning said, "Both language understanding and artificial intelligence require being able to understand bigger things from knowing about *smaller parts*."[9] Firstly, it is more reasonable to build a bottom-up system with linguistically structured representations like syntax or semantic parsers and sub-word structures (refer to Sect. 4.2) than an end-2-end system without consideration of any linguistic structures. Moreover, we can use some constrains to normalize each subcomponent and make it understandable for humans, as well as relieve the non-convergent problems. For instance, an attention mechanism [3] is one of the most successful mechanisms from the point view of normalization. For an unconstrained real-valued vector, it is hard to understand and know how it works. After the addition of a softmax operation, this vector denotes a multinomial probability distribution in which each element ranges from 0 to 1 and the sum of the vectors equals 1.

Contextualized Word Embedding Previously, word embedding was static, which means it did not depend on the context and it was one-to-one mapping from a word to a static vector. For example, the word "bank" has at least two meanings, i.e., "the land alongside or sloping down to a river or lake" and "a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency." However, the word in a finance-related context and a river-related context could be mapped into the same fixed vector, which is not reasonable for language. Instead of storing a static look-up table, contextualized word embedding learns a language model to generate a real-time word vector for each word based on the neighboring word (context). The first model was proposed with the name Embedding from Language MOdel (ELMO) [80], and it was further investigated by Generative Pre-Training (GPT) [83] and BERT [25]. More specifically, BERT obtained new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE benchmark, MultiNLI accuracy, and the SQuAD with huge improvements.

Linguistically Enhanced Word Embedding One of the main criticisms of word embedding is that it ignores the linguistic knowledge and instead adopts a brute force approach which is totally driven by data. However, there are already many

[9] https://nlp.stanford.edu/manning/talks/Simons-Institute-Manning-2017.pdf.

linguistic resources for words, e.g., WordNet and sentimental lexicon. Incorporation of the linguistic knowledge trends in the current paradigm of the NLP can relieve the dependence of data. These linguistic resources are expected to enhance the representative power of word embedding, which may be used in higher layers than word embedding layers like syntax structures [101] or only word embedding with WordNet and related lexicon resources [30, 61].

Sub-Word Embedding We briefly discussed the OOV problem in Sect. 4.1. Previous solutions for relieving it were commonly based on empirical insights, e.g., assigning a special token to all the OOV words. In [132] character-based embedding for text classification was adopted, avoiding directly processing the word-level embedding. In the sub-word embedding, there are no OOV problems since the proposed approaches directly build the word embedding with the units with smaller granularity which may have a limited number. For example, one of the sub-word approaches in English is based on characters, which are limited to a–z, A–Z, 0–9, punctuation, and other special symbols. Moreover, a character level approach could be beneficial for some specific languages, like Chinese, that can make use of smaller-granularity units which are smaller than words but also have abundant semantic information, like components. Sub-word regularization [54] trains the model with multiple sub-word segmentation (based on a unigram language model) probabilistically sampled during training. These works demonstrate that there is some potential to incorporate some fine-refined linguistic knowledge in the neural network [13, 54].

Advanced Word Embedding: Beyond one Fixed Real-Valued Vector More recently, different types of word embedding beyond real-valued vectors have been developed, for example:

- **Gaussian embedding** [112] assigns a Gaussian distribution for each word, instead of a point vector in a low-dimension space. The advantages are that it naturally captures uncertainty and expresses asymmetries in the relationship between two words.
- **Hyperbolic embedding** [77, 93] embeds words as points in a Cartesian product of hyperbolic spaces; therefore, the hyperbolic distance between two points becomes the Fisher distance between the corresponding probability distribution functions (PDFs). This additionally derives a novel principled "is-a" score on top of word embeddings that can be leveraged for hypernymy detection.
- **Meta embedding** [50, 127] adopts multiple groups of word vectors and adaptively obtains a word vector by leveraging all the word embeddings.
- **Complex-valued embedding** [62, 114] formulates a linguistic unit as a complex-valued vector, and links its length and direction to different physical meanings: the length represents the relative weight of the word, while the direction is viewed as a superposition state. The superposition state is further represented in an amplitude-phase manner, with amplitudes corresponding to the lexical meaning and phases implicitly reflecting the higher-level semantic aspects such as polarity, ambiguity, or emotion.

4.3 Linking Word Embedding to Vector-Space Based Approaches and Representation of Thematic Structures

Deriving the Topic Distribution from Word Embedding Research on the representation of themes in an unstructured document corpus—finding word patterns in a document collection—dates back to the 1990s, i.e., to the introduction of LSA [24]. A subsequent extension that exploits a statistical model was proposed By Hofmann in[44]. That model, named Probabilistic Latent Semantic Indexing (PLSI), relies on the *aspect model*, a latent variable model for co-occurrence data where an occurrence—in our case a word occurrence—is associated with an unobserved/latent variable. The work by Hofmann and subsequent works rely on the "same fundamental idea—that a document is a mixture of topics—but make slightly different statistical assumptions" [99]. For instance, in [12] Blei et al. extended the work by Hofmann making an assumption on how the mixture weights for the topics in a document are generated, introducing a Dirichlet prior. This line of research is known as *topic modeling*, where a topic is interpreted as a group of semantically related words. Since the focus of this paper is not on topic modeling, in the remainder of this section we are going to introduce only the basic concepts needed to discuss possible links with word embedding approaches; the reader can refer to the work reported in [9, 11, 15, 99] for a more comprehensive discussion on the difference among the diverse topic models and the research trends and direction in topic modeling.

As mentioned above, probabilistic topic models consider the document as a distribution over topics, while the topic is a distribution over words. In PLSI no prior distributions are adopted and the joint probability distribution between document and word is expressed as follows:

$$p(w,d) = \sum_{c \in C} p(w,d,c) = p(d) \sum_{c \in C} p(c,w|d) = p(d) \sum_{c \in C} p(c|d)p(w|c), \tag{1}$$

where d is a document, while w is a specific word and C is the collection of topics. A crucial point of topic models is how to estimate the $p(c|d)$ and $p(w|c)$.

Using an "empirical" approach, we can also get the $p(c|d)$ and $p(w|c)$ from word embedding. Suppose that we obtain a word embedding, i.e., a mapping from a word (denoted as an index with a natural number) to a dense vector $\mathcal{N} \to \mathcal{R}^n$. For a given sentence S with words sequence $\{w_1, w_2, \dots w_n\}$, we can get a representation for s with an average embedding like [49], namely $\boldsymbol{d} = \sum_{i=1}^{n} \boldsymbol{w_i}$. It is easy to define a topic with distribution $p(w|c)$, represented as: $\boldsymbol{c_j} = \sum_{i=1}^{|V|} p_{w_i|c} \boldsymbol{w_i}$, $\boldsymbol{c_j} \in C$. Then we can obtain the following topic distribution of a document:

$$p(c_j|d) = \frac{e^{-||\boldsymbol{d}-\boldsymbol{c_j}||_2}}{\sum_i^{|C|} e^{-||\boldsymbol{d}-\boldsymbol{c_i}||_2}}. \tag{2}$$

The relationship between word embedding and topic models has been addressed in the literature. For instance, the work reported in [60] shows that a special case of word embedding, i.e., Skip-gram, has the same optimal solution as the factorization of a shifted Point-wise Mutual Information (PMI) matrix.[10] Empirically, the count-based representations and distributed representations can be combined together with complementary benefits [78, 115].

Recent works focused on exploiting both methods. The discussion of previous approaches reported in [98] reports on two lines of research: methods used to improve word embedding through the adoption of topic models, which addresses the polysemy problem; methods used to improve topic models through word embedding, which obtains more coherent words among the top words associated with a topic. These approaches mainly rely on a pipeline strategy, "where either a standard word embedding is used to improve a topic model or a standard topic model is used to learn better word embeddings" [98]. The limitation of these approaches is the lack of capability to exploit the mutual strengthening between the two, which a joint learning strategy, in principle, could exploit. This is the basic intuition underlying the work reported in [98]. Another example is *lda2vec* where the basic idea was "modifying the Skip-gram Negative-Sampling objective in [71] to utilize document-wide feature vectors while simultaneously learning continuous document weights loading onto topic vectors." The work reported in [117] proposes a different approach relying on a "topic-aware convolutional architecture" and a reinforcement learning algorithm in order to address the task of text summarization.

Regarding the Contextual Windows The previous subsection suggests possible connections between word embedding and the representation of the thematic structure in document corpora, e.g., through topic models. Vector space based approaches in IR, topic models, matrix factorization, and word embedding can be considered as different approaches relying on distributional hypothesis as discussed in Sect. 1. One of the differences among these methods may be how to choose the size of the contextual window. In this paper, we classify the contextual window into several sizes, i.e., "character → word → phase/N-gram → clause → sentence → paragraph → document," ordered from the smallest to the biggest granularity. For example, VSM in IR usually chooses the whole document as the context; thus, it may capture the document-level feature of text, like the thematic structure. Approaches based on word-word matrix factorization usually set a smaller window size to statistically analyze the co-occurrence between words—similar to the windows of CBOW [70], thus targeting a smaller context in order to capture the word-level feature related to its word meaning.

Depending on the context size, features in vector space based approaches in IR are already at a relatively high level, e.g., the TFIDF vector or the language model [130], and they can be used directly for relatively downstream task like

[10]The shifted PMI matrix is "the well-known word-context PMI matrix from the word-similarity literature, shifted by a constant offset" [60].

document ranking. Lower-level word features of word-word matrix factorization (or CBOW) can be used directly for the relatively upstream task like morphology, lexicon, and syntax, and it needs some abstraction components to extract from the low-level features to high-level features. On the other hand, abstraction from the low-level features to high-level features may imply a loss of some fundamental lexical meaning. The low-level features (word-word matrix factorization or CBOW) are usually considered a better basic input for another "stronger" learning model—e.g., when using multiple layers of non-linear abstraction—compared to higher-level features.

4.4 Towards Dynamic Word Embedding

One of the limitations of most representations of words, documents, and themes is that they do not consider the *temporal dimension*. This is crucial when considering corpora such as historical document archives, newspapers, or social media, e.g., tweets, that consist in a continuous stream of informative resources. The use of these "time-stamped" resources is useful not only for general tasks but also for specialist users. Indeed, the tasks performed by the specialists of a discipline need to make hypotheses from data, for example, by means of longitudinal studies. This is the case for the tasks performed by specialists in the field of Social Science, Humanities, Journalism, and Marketing.

Let us consider, for instance, the case of sociologists that study the public perception of science and technology by the public opinion—this line of research is known as STS, Science and Technology Studies. The study of how some science and technology-related issues are discussed by the media, e.g., newspapers, could be useful in providing policy makers with insights on the public perception of some issues on which they should or intend to take actions or provide guidance on the way these issues should be publicly discussed (e.g., on the use of "sensible" words or aspects related to the issues). In this context, relevant information can be gained from how the meaning of a word or how the perception of an issue related to a word change through time.

Previous works on topic modeling addressed the issue of including the temporal dimension, specifically, the issue that topics can change over time. In [72] Mimno proposes a possible approach to visualize the topic coverage across time starting from topic learnt using a "static" approach: given the probabilities and the topic assignment estimated via LDA, the topic trend can be visualized by counting the number of words in each topic published in a given year and then normalizing over the total number of words for that year. Other works embedded the time dependence directly in the statistical model. One of the earliest works is that proposed in [10] where dynamic topic models were introduced. The underlying assumption is that time is divided into time slices, e.g., by years; documents in a specific time slice are modeled using a K-component topic model—K is the number of topics—where topics in a given time slice evolve from those in the previous time slice. This kind of representation could be extremely useful for a specialist in order to follow the evolution of a single word, e.g., by inspecting the top words for diverse topics where the word is framed in his research hypothesis—e.g., the "nuclear" word framed in "innovation," "risk," or "energy" topics—or following the posterior estimate of the frequency of the word as a function of the year, as shown in [10]. As stated by the authors, one of the limitations of that approach is that the number of topics needs to be specified beforehand; the work reported in [28] aimed to address this limitation by introducing a non-parametric version for modeling topics over time.

Even if dynamic/time-aware versions of topic models can support specialists in their investigation, the adoption of word embedding to study changes in a word representation could provide complementary evidence to support or undermine a research hypothesis. Indeed, as mentioned above, topic models are learned from a more "global view," while word embedding exploits a more "local view," e.g., using evidence from local context windows; this local view might help to obtain a word representation that, in a way, "reflects the semantic, and sometimes also syntactic, relationships between the words" [98]. Another point of view about the difference between topic models and word embedding approaches could be the scale of the dimension and the sparseness degree in the vector space. Intuitively, topic models (especially the topic distribution over words) tend to adopt sparse vectors with bigger dimensions, while the word embedding approaches adopt low-dimension dense vectors which may save some memory space and provide more flexibility for the high-level applications. Note that the difference in sparseness can be decreased to some extent by the sparsing regularization as introduced by Vorontsov et al. [113].

The work reported in [55] discussed several approaches to identify "linguistic change." As an example of linguistic change, they referred to the change of the word "gay" that shifted from the meaning of "cheerful" or "frolicsome" to homosexuality (see Fig. 1 of that paper). They proposed three different approaches to generate time series aimed to capture different aspects of word evolution across time: a frequency-based method, a syntactic method, and a distributional method. Because of the objective of this survey, we will focus on the last one. They divided the entire time span of the dataset in time slices of the same size, e.g., 1-month or 5-year slices. Then a word embedding technique—`gensim` implementation of the Skip-gram model—was used to learn word representation in each slice; an alignment procedure was then adopted to consider all the embeddings in a unique coordinate system. Finally, the time series was obtained by calculating the distance between the time 0 and the time t in the embedding space of the final time slice. The use of time series has several benefits, e.g., the possibility to use change point detection methods to identify the point in time where the new word meaning became predominant. The distributional approach was the most effective in the various evaluation settings: synthetic evaluation, evaluation on a reference dataset, and evaluation with human assessors.

In [37] the change in meaning of a word through time is referred to as a "semantic change." The authors report several examples in word meaning change, e.g., the semantic change of the word "gay" as in [55] and that of the word "broadcast," which at the present time is mainly intended as a synonym of "transmitting signal."

While in the early twentieth century it meant "casting out seeds." In that work, static versions of word embedding techniques were used, but word embedding was learned for each time slice and then aligned in order to make word vectors from different time periods comparable; aligning is addressed as an Orthogonal Procrustes Problem. Three word embedding techniques were considered. The first is based on Positive Point-wise Mutual Information (PPMI) representations, where PPMI values are computed with respect to pre-specified context words and are prepared in a matrix whose rows are the word vector representations. The second approach, in the paper referred to as SVD, considers a truncated version of the SVD of the PPMI matrix. The last method is Skip-gram with negative sampling. The work reported in that paper is pertinent to our "specialist user scenario" since the main contribution is actually a methodology to investigate two research hypotheses. In particular, the second hypothesis investigated is that "Polysemous words change at faster rates"; this is related to an old hypothesis in linguistics that dates back to [16] and states that "words become semantically extended by being used in diverse contexts." Subsequent works [29] show that the results obtained in the literature for diverse hypotheses on semantic change—including those in [37]—should be revised; using as a control test an artificially generated corpus with "no semantic change" as a control test, they showed that the previously proposed methodologies detected a semantic change in the control test as well. The same result was observed for diverse hypotheses—see the survey reported in [56] for an overview of the diverse hypotheses investigated. As mentioned by Dubossarsky et al. [29], their result supports further research in evaluation of dynamic approaches "articulating more stringent standards of proof and devising replicable control conditions for future research on language change based on distributional semantics representations" [29].

The work reported in [90] introduces a dynamic version of the exponential family of embedding previously proposed in [91]. The reason for the introduction of the exponential family of embedding was to generalize the idea of word embedding to other data, e.g., neuronal activity or shopping for an item on the basis of the context (other items in the shopping cart). The obtained results show that the dynamic version of the exponential family embedding provides better results in terms of conditional likelihood of held-out predictions when compared with static embeddings [71, 91] and time-binned embeddings [37].

In [4] the authors extend the Bayesian Skip-gram Model proposed in [5] to a dynamic version considering a diffusion process of the embedding vectors over time, more specifically a Ornstein–Uhlenbeck process. Both of the two proposed variants resulted in more smoothed word embedding trajectories[11] than the baselines, which utilized the approach proposed in [37].

In [125] the authors proposed to find temporal word embedding to solve a joint optimization problem where the "key" component is a smoothing term that

[11]Trajectories where based on the cosine distance between two words representation over time.

encourages embedding to be aligned, thus explicitly solving the alignment problem while learning embedding and avoiding a two-step strategy like that adopted in [37] or in [55].

In [56] the authors report a number of open issues concerning the study of temporal aspects of semantic shifts. Two challenges that are particularly relevant to the works reported in this chapter and this venue are: (1) *the lack of formal mathematical models of diachronic embeddings*; (2) the need for *robust gold standard test sets of semantic shifts*; (3) the need for *algorithms able to work on small datasets*. With regard to the first point, investigating quantum-inspired models could be a possible research direction to find a formal mathematical framework to model dynamic/diachronic word embeddings, e.g., exploiting the generalized view of probability and the theory of time evolution of systems. With regard to the second point, and evaluation in general, a possible direction is to devise tasks with specialists, e.g., journalists, linguists, or social scientists, to create adequate datasets. This is also related to the last point, i.e., the need for algorithms that are "robust" to the size of the dataset: indeed, specialists, even when performing longitudinal user studies, can rely on relatively small datasets in order to investigate specific research issues. On the basis of the ongoing collaboration with sociologists and linguists, another open issue that could be really beneficial for the specialists investigations is "identifying groups of words that shift together in correlated ways" [56]; this could be particularly useful to investigate how some thematic issues are perceived by the public opinion and how this perception varies through time. As suggested by the results reported in [29], evaluation protocols to measure these algorithms' effectiveness should be rigorously designed.

As mentioned above, word embedding and topic models are based on two very different views. Rudolph et al. [90] suggest another possible research direction in the dynamic representation of words: devise models able to combine the two approaches and exploit their "complementary" representations in dynamic settings.

5 Conclusion

We introduced many vector space based approaches for representing words, especially the word vector techniques. Regarding the word vector, we introduced many variants presented throughout in the history and their limitations and trends. A concise summary is reported in Table 4.

Since the effectiveness of word embedding is supported by the investigation in many NLP and IR tasks and by many benchmarks, it is worth investigating further. In the future, it is expected to incorporate some external knowledge like linguistic features or the common sense of humans (like knowledge base) to word vectors. Besides these empirical efforts, some theoretical understanding is also important to this field, like the interpretability about why it works and where it does not work.

Table 4 A summary including various word vector techniques

Algorithm	Polysemy	Interpretability	OOV	Speed
NNLM [7]				
C&W [21]				
Skip-gram [70]				+
CBOW [70]				+
Glove [78]		+		
Char-based embedding [132]			+	
Elmo [80]	+			
BERT [25]	+			
Gaussian embedding [112]	+			
Hyperbolic embedding [77]		+		
Meta embedding [127]	+			
Complex embedding [62]		+		

Some earlier works aim to develop fast-training methods, while recent works focus more on the empirical performance with dynamic context-aware embedding and the intuitive understanding of interpretable word vectors

References

1. Aggarwal, C. C., & Zhai, C.-X. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163–222). Berlin: Springer.
2. Athiwaratkun, B., & Wilson, A. G. (2017). Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1645–1656).
3. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Preprint. arXiv:1409.0473.
4. Bamler, R., & Mandt, S. (2017). Dynamic word embeddings. Preprint. arXiv:1702.08359.
5. Barkan, O. (2017). Bayesian neural word embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 3135–3143).
6. Barkan, O., & Koenigstein, N. (2016). Item2Vec: Neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). Piscataway: IEEE.
7. Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research, 3*, 1137–1155.
8. Bian, W., Li, S., Yang, Z., Chen, G., & Lin, Z. (2017). A compare-aggregate model with dynamic-clip attention for answer selection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1987–1990). New York: ACM.
9. Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77 (2012).
10. Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06* (pp. 113–120). New York: ACM.
11. Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Text mining: classification, clustering, and applications*. *Data mining and knowledge discovery series*. London: Chapman & Hall.
12. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
13. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

14. Bowman, S. R., Angeli, G., Potts, C., & Manning C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg: Association for Computational Linguistics.
15. Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval, 11*(2–3), 143–296.
16. Bréal, M. (1897). *Essai de Sémantique: Science des significations*. Paris: Hachette.
17. Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467–479.
18. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. Preprint. arXiv:1409.1259.
19. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, Schwenk, F. H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Preprint. arXiv:1406.1078.
20. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-T., Choi, Y., et al. (2018). QuAC: Question answering in context. Preprint. arXiv:1808.07036.
21. Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167). New York: ACM.
22. Cui, H., Sun, R., Li, K., Kan, M.-Y., & Chua, T.-S. (2005). Question answering passage retrieval using dependency relations. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 400–407). New York: ACM.
23. Cui, P., Wang, X., Pei, J., & Zhu, W. (2018). A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering, 31*(5), 833–852.
24. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.
25. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint. arXiv:1810.04805.
26. Ding, Y., Liu, Y., Luan, H., & Sun, M. (2017). Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (vol. 1, pp. 1150–1159).
27. Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005).*
28. Dubey, A., Hefny, A., Williamson, S., & Xing, E. P. (2013). A nonparametric mixture model for topic modeling over time. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 530–538).
29. Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (vol. 1, pp. 1136–1145). Stroudsburg: Association for Computational Linguistics.
30. Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. Preprint. arXiv:1411.4166.
31. Fellbaum, C. (2000). WordNet: An electronic lexical database. *Language, 76*(3), 706.
32. Feng, M., Xiang, B., Glass, M. R., Wang, L., & Zhou, B. (2015). Applying deep learning to answer selection: A study and an open task. Preprint. arXiv:1508.01585.
33. Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis (special volume of the Philological Society)* (vol. 1952–59, pp. 1–32). Oxford: The Philological Society.
34. Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. Preprint. arXiv:1705.03122.

35. Gittens, A., Achlioptas, D., & Mahoney, M. W. (2017). Skip-gram-zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (vol. 1, pp. 69–76).
36. Goller, C., & Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. *Neural Networks, 1*, 347–352.
37. Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. Preprint. arXiv:1605.09096.
38. Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162.
39. He, H., Gimpel, K., & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1576–1586).
40. He, H., & Lin, J. (2016). Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 937–948).
41. He, K., Girshick, R., & Dollár, P. (2018). Rethinking ImageNet pre-training. Preprint. arXiv:1811.08883.
42. Heilman, M., & Smith, N. A. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 1011–1019). Stroudsburg: Association for Computational Linguistics.
43. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
44. Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99* (pp. 50–57). New York: ACM.
45. Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems* (pp. 2042–2050).
46. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. Preprint. arXiv:1508.01991.
47. Iyyer, M., Yih, W.-T., & Chang, M.-W. (2017). Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (vol. 1, pp. 1821–1831).
48. Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. Preprint. arXiv:1705.03551.
49. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 427–431). Stroudsburg: Association for Computational Linguistics.
50. Kiela, D., Wang, C., & Cho, K. (2018). Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1466–1477).
51. Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751).
52. Kim, Y. (2014). Convolutional neural networks for sentence classification. Preprint. arXiv:1408.5882.
53. Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., et al. (2018). The narrativeQA reading comprehension challenge. *Transactions of the Association of Computational Linguistics, 6*, 317–328.
54. Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

55. Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 625–635).
56. Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Stroudsburg: Association for Computational Linguistics.
57. Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association of Computational Linguistics* (to appear). https://tomkwiat.users.x20web.corp.google.com/papers/natural-questions/main-1455-kwiatkowski.pdf
58. Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems, 31*(6), 5–14.
59. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. Preprint. arXiv:1603.01360.
60. Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems* (pp. 2177–2185).
61. Li, J., Hu, R., Liu, X., Tiwari, P., Pandey, H. M., Chen, W., et al. (2019). A distant supervision method based on paradigmatic relations for learning word embeddings. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-019-04071-6
62. Li, Q., Uprety, S., Wang, B., & Song, D. (2018). Quantum-inspired complex word embedding. In *Proceedings of the Third Workshop on Representation Learning for NLP, Melbourne* (pp. 50–57). Stroudsburg: Association for Computational Linguistics.
63. Lin, D., & Wu, X. (2009). Phrase clustering for discriminative learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2* (pp. 1030–1038). Stroudsburg: Association for Computational Linguistics.
64. Lin, R., Liu, S., Yang, M., Li, M., Zhou, M., & Li, S. (2015). Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 899–907).
65. Lipton, Z. C. (2016). The mythos of model interpretability. Preprint. arXiv:1606.03490.
66. Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. Preprint. arXiv:1705.11168.
67. Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers, 28*(2), 203–208.
68. Melucci, M. (2015). *Introduction to information retrieval and quantum mechanics*. Berlin: Springer.
69. Melucci, M., & van Rijsbergen, C. J. (2011). *Quantum mechanics and information retrieval* (chap. 6, pp. 125–155). Berlin: Springer.
70. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Preprint. arXiv:1301.3781.
71. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).
72. Mimno, D. (2012). Computational historiography. *Journal on Computing and Cultural Heritage, 5*(1), 1–19.
73. Mitra, B., & Craswell, N. (2017). Neural models for information retrieval. Preprint. arXiv:1705.01509.
74. Mitra, B., & Craswell, N. (2018). An introduction to neural information retrieval.*Foundations and Trends in Information Retrieval, 13*(1), 1–126.
75. Mnih, A., & Hinton, G. (2007). Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 641–648). New York: ACM.

76. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., et al. (2016). MS MARCO: A human generated machine reading comprehension dataset. Preprint. arXiv:1611.09268.
77. Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems* (pp. 6338–6347).
78. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (vol. 14, pp. 1532–1543).
79. Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* (pp. 183–190). Stroudsburg: Association for Computational Linguistics.
80. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. Preprint. arXiv:1802.05365.
81. Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence, 46*(1–2), 77–105.
82. Punyakanok, V., Roth, D., & Yih, W.-T. (2004). Mapping dependencies trees: An application to question answering. In *Proceedings of AI&Math 2004* (pp. 1–10).
83. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\understanding_paper.pdf
84. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. Preprint. arXiv:1806.03822.
85. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. Preprint. arXiv:1606.05250.
86. Rao, J., He, H., & Lin, J. (2016). Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 1913–1916). New York: ACM.
87. Reddy, S., Chen, D., & Manning, C. D. (2018). CoQA: A conversational question answering challenge. Preprint. arXiv:1808.07042.
88. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333–389.
89. Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323–2326.
90. Rudolph, M., & Blei, D. (2018). Dynamic Bernoulli embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1003–1011).
91. Rudolph, M., Ruiz, F., Mandt, S., & Blei, D. (2016). Exponential family embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16* (pp. 478–486). Red Hook: Curran Associates Inc.
92. Saha, A., Pahuja, V., Khapra, M. M., Sankaranarayanan, K., & Chandar, S. (2018). Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. Preprint. arXiv:1801.10314.
93. Sala, F., De Sa, C., Gu, A., & Ré, C. (2018). Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning* (pp. 4457–4466).
94. Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 298–307).
95. Severyn, A., & Moschitti, A. (2013). Automatic feature engineering for answer selection and extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 458–467).
96. Severyn, A., & Moschitti, A. (2015). Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 373–382). New York: ACM.
97. Shen, G., Yang, Y., & Deng, Z.-H. (2017). Inter-weighted alignment network for sentence pair modeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1179–1189).

98. Shi, B., Lam, W., Jameel, S., Schockaert, S., & Lai, K. P. (2017). Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17* (pp. 375–384). New York: ACM.
99. Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis* (pp. 424–440).
100. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* (pp. 3104–3112).
101. Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. Preprint. arXiv:1503.00075.
102. Talmor, A., & Berant, J. (2018). The web as a knowledge-base for answering complex questions. Preprint. arXiv:1803.06643.
103. Tay, Y., Luu, A. T., & Hui, S. C. (2017). Enabling efficient question answer retrieval via hyperbolic neural networks. CoRR abs/1707.07847.
104. Tay, Y., Phan, M. C., Tuan, L. A., & Hui, S. C. (2017). Learning to rank question answer pairs with holographic dual LSTM architecture. Preprint. arXiv:1707.06372.
105. Tay, Y., Tuan, L. A., & Hui, S. C. (2018). Multi-cast attention networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2299–2308). New York: ACM.
106. Tran, Q. H., Lai, T., Haffari, G., Zukerman, I., Bui, T., & Bui, H. (2018). The context-dependent additive recurrent neural net. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (vol. 1, pp. 1274–1283).
107. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2016). NewsQA: A machine comprehension dataset. Preprint. arXiv:1611.09830.
108. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., et al. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP* (pp. 191–200). Stroudsburg: Association for Computational Linguistics.
109. Upadhyay, S., Chang, K. W., Taddy, M., Kalai, A., & Zou, J. (2017). Beyond bilingual: Multi-sense word embeddings using multilingual context. Preprint. arXiv:1706.08160.
110. Van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge: Cambridge University Press.
111. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
112. Vilnis, L., & McCallum, A. (2014). Word representations via Gaussian embedding. Preprint. arXiv:1412.6623.
113. Vorontsov, K., Potapenko, A., & Plavin, A. (2015). Additive regularization of topic models for topic selection and sparse factorization. In *International Symposium on Statistical Learning and Data Sciences* (pp. 193–202). Berlin: Springer.
114. Wang, B., Li, Q., Melucci, M., & Song, D. (2019). Semantic Hilbert space for text representation learning. Preprint. arXiv:1902.09802.
115. Wang, B., Niu, J., Ma, L., Zhang, Y., Zhang, L., Li, J., et al. (2016). A Chinese question answering approach integrating count-based and embedding-based features. In *Natural Language Understanding and Intelligent Applications* (pp. 934–941). Cham: Springer.
116. Wang, D., & Nyberg, E. (2015). A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (vol. 2, pp. 707–712).
117. Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., & Du, Q. (2018). A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. Preprint. arXiv:1805.03616.

118. Wang, M., & Manning, C. D. (2010). Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1164–1172). Stroudsburg: Association for Computational Linguistics.
119. Wang, M., Smith, N. A., & Mitamura, T. (2007). What is the jeopardy model? A quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
120. Wang, Z., Hamza, W., & Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. Preprint. arXiv:1702.03814.
121. Wong, S. K. M., Ziarko, W., & Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '85* (pp. 18–25). New York: ACM.
122. Yang, L., Ai, Q., Guo, J., & Croft, W. B. (2016). aNMM: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 287–296). New York: ACM.
123. Yang, Y., Yih, W.-T., & Meek, C. (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2013–2018).
124. Yao, X., Van Durme, B., Callison-Burch, C., & Clark, P. (2013). Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 858–867).
125. Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18* (pp. 673–681). New York: ACM.
126. Yih, W. T., Chang, M. W., Meek, C., & Pastusiak, A. (2013). Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (vol. 1, pp. 1744–1753).
127. Yin, W., & Schütze, H. (2015). Learning meta-embeddings by using ensembles of embedding sets. Preprint. arXiv:1508.04257.
128. Yu, L., Hermann, K. M., Blunsom, P., & Pulman, S. (2014). Deep learning for answer sentence selection. Preprint. arXiv:1412.1632.
129. Zamani, H., & Croft, W. B. (2017). Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'17* (pp. 505–514). New York: ACM.
130. Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *Transactions on Information Systems, 22*(2), 179–214.
131. Zhang, P., Niu, J., Su, Z., Wang, B., Ma, L., & Song, D. (2018). End-to-end quantum-like language models with application to question answering. In *The Thirty-Second AAAI Conference on Artificial Intelligence*. Menlo Park: Association for the Advancement of Artificial Intelligence.
132. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657). Cambridge: MIT Press.

Quantum-Based Modelling of Database States

Ingo Schmitt, Günther Wirsching, and Matthias Wolff

Abstract Database design of real-world scenarios requires complex data structures in order to adequately model complex real-world objects. Complex data structures can be constructed by a recursive use of elementary data types and data type constructors. The mathematics behind quantum mechanics provides us an interesting theory combining concepts from linear algebra, probability calculus, and logic. In order to make the mathematics of quantum mechanics available for database structures and states we develop a mapping of concepts from type theory of databases to the mathematics of quantum mechanics.

1 Introduction

The mathematics behind quantum mechanics [1] provides us a formalism that combines very elegantly concepts from probability calculus, linear algebra, and logic. The semantics of a quantum system is expressed by a normalized ket vector in an inner product space. Here we show how to model complex data structures of a database state as a normalized ket vector of an inner product space, see also [2]. Furthermore, we show how to read a database vector by use of the statistics of quantum measurement. Our database mapping to the mathematics of quantum mechanics proposed in the following is restricted to finite dimensional and real inner product spaces. For a query language based on our mapping and quantum logic as well as quantum measurement, we refer to [3].

Please note that we do not propose to perform a mapping onto a physical quantum computer. Instead, the proposed mapping is on a conceptual level rather

I. Schmitt (✉) · M. Wolff
Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany
e-mail: schmitt@b-tu.de

G. Wirsching
Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany
e-mail: gunther.wirsching@ku.de

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_6

than on an implementation level. The benefit from doing so is to bridge well-known database modelling concepts into another formalism. This is very promising because the rich theory of linear algebra and quantum logic [4] provides us powerful concepts and gives us a deep understanding of certain database problems. For example, the relation between entanglement and functional dependencies between database values, and reasoning from databases based on quantum logic are not well understood so far. On a conceptual level we are able to develop and to prove interesting new theorems.

2 Motivating Example: Car Dealership

In the following we develop an example for demonstrating a quantum-based data modelling of a database structure and the measurement of its state. As example we use the managed data objects of a car dealership. From the view of a database designer, cars are complex-structured objects. Every car is composed of different technical components as shown in Fig. 1. Furthermore, a service booklet containing a record of car inspections exists for every car.

Some properties of components of the car dealership are listed in Table 1 defining the state of a car. Furthermore, some atomic conditions for measurements based on these properties are given in Table 2.

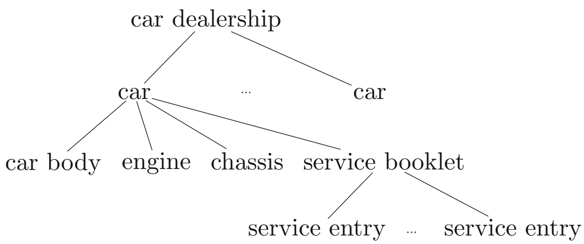


Fig. 1 Components of the car management

Table 1 Properties of car components

Component	Property	Value domain
Car	`license tag`	Set of valid license tags
Car	`year of construction`	2000–2020
Engine	`number of cylinders`	2–16
Engine	`cylinder arrangement`	Row, v-form, Boxer-form
Engine	`fuel tank` (l)	30–80
Car body	`kilometre` (km)	0–300.000
Car body	`shipping volume` (l)	200–500
Service entry	`date`	01.01.2000 to 31.12.2030
Service entry	`kilometre` (km)	0–300.000

Table 2 Atomic conditions on car properties

Label	Condition
`YC1`	`year of construction` = 2016
`YC2`	`year of construction` = 2017
`FT1`	`fuel tank` ≈ 35
`FT2`	`fuel tank` is very large
`K1`	`kilometre` ≈ 15.000
`K2`	`kilometre` is very small
`NC`	`number of cylinders` = 4
`CA1`	`cylinder arrangement` = Row
`CA2`	`cylinder arrangement` = Boxer

When we look at condition `FT2` we make the following observation: testing `FT2` against the state of a car object cannot adequately return yes or no. Instead, we expect to receive a grade of compliance from the interval [0, 1]. A high value signals a strong compliance and vice versa. Later on, we will show how the statistics of quantum measurements provides us a mean to compute the required gradual values.

At first, we discuss how to model elementary data types by using the mathematics behind quantum mechanics. Here we focus on finite dimensional and real inner product spaces. Later on, we will explain how to construct complex data types and how to map them into the quantum world.

3 Modelling Elementary Data Types

An elementary data type defines a data structure and operations to deal with its values. A data type is elementary if its values cannot be meaningfully decomposed into smaller semantic values. In our example the property `year of construction` is elementary. Its domain covers all possible year values of car construction. A useful operation could be the computation of the difference between 2 year values. We define the function *dom* which assigns to a data type a set of valid values. That set is often called *domain* of a data type.

We distinguish between two types of elementary data types:

- *orthogonal data type*: The values of that data type are independent from each other. There is no meaningful similarity between them. Two values are either identical or not identical. In our example, the property `cylinder arrangement` is orthogonal.
- *non-orthogonal data type*: Besides the test on identity between two values gradual similarity values can be required between them. In our example the property `fuel tank` is non-orthogonal: a required volume of 35 L is more similar to a given value of 40 L than 45 L.

The distinction between orthogonal and non-orthogonal often depends on the intended application semantics. In some application it may be important to demand for an exact value of 35 L for a fuel tank and every deviation is seen as wrong. In that case, `fuel tank` would be modelled as an orthogonal data type. For simplicity, in the following we assume that every property is categorized either as orthogonal or non-orthogonal.

In next subsections we show how to map an elementary data type $\langle \mathtt{dt} \rangle$ with a finite domain

$$Dom\,(\langle \mathtt{dt} \rangle) := \{\mathrm{V}_1, \ldots, \mathrm{V}_k\}$$

to a family of ket vectors of an inner product space. The mapping of a value to a ket vector is denoted by the symbol $\mapsto$. Function $QDom$ assigns to a data type the set of ket vectors which appear as possible outcome of this mapping.

3.1 Orthogonal Data Types

The values of an orthogonal data type $\langle \mathtt{dt} \rangle$ are bijectively mapped to ket vectors forming an orthonormal basis of an inner product space:

$$\begin{aligned} QDom(\langle \mathtt{dt} \rangle) &= \{|\mathrm{V}_1\rangle, \ldots, |\mathrm{V}_k\rangle\} \\ Dom(\langle \mathtt{dt} \rangle) &\rightarrow QDom(\langle \mathtt{dt} \rangle) \\ \forall i \in [1, k] : \mathrm{V}_i &\mapsto |\mathrm{V}_i\rangle. \end{aligned}$$

The corresponding ket vectors are taken to be mutually orthogonal; they span a k-dimensional inner product space.

Let us take a basis ket vector $|\mathrm{V}_x\rangle$ for a value of an orthogonal property. If we want to test the value V_i for identity with V_x, we proceed in a way reflecting quantum measurement. We construct the projector $P = |\mathrm{V}_i\rangle\langle \mathrm{V}_i|$:

$$\langle \mathrm{V}_x|P|\mathrm{V}_x\rangle = \langle \mathrm{V}_x|\mathrm{V}_i\rangle\langle \mathrm{V}_i|\mathrm{V}_x\rangle = \begin{cases} 1 \text{ if } i = x \\ 0 \text{ otherwise.} \end{cases}$$

For testing whether a value x is contained in a value set $S = \{s\}$ we use the projector $P = \sum_{s \in S} |\mathrm{V}_s\rangle\langle \mathrm{V}_s|$:

$$\begin{aligned} \langle \mathrm{V}_x|P|\mathrm{V}_x\rangle &= \langle \mathrm{V}_x| \left(\sum_{s \in S} |\mathrm{V}_s\rangle\langle \mathrm{V}_s| \right) |\mathrm{V}_x\rangle \\ &= \sum_{s \in S} \langle \mathrm{V}_x|\mathrm{V}_s\rangle\langle \mathrm{V}_s|\mathrm{V}_x\rangle = \begin{cases} 1 \text{ if } x \in S \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

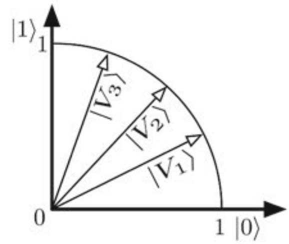


Fig. 2 Value mapping into a real one-qubit system

$$|\mathrm{V}_1\rangle = 0.9 \cdot |0\rangle + 0.435 \cdot |1\rangle$$
$$|\mathrm{V}_2\rangle = 0.7 \cdot |0\rangle + 0.714 \cdot |1\rangle$$
$$|\mathrm{V}_3\rangle = 0.3 \cdot |0\rangle + 0.954 \cdot |1\rangle$$

3.2 Non-orthogonal Data Types

Between values of a non-orthogonal data type ⟨`dt`⟩ a gradual similarity is required. Therefore we choose non-orthogonal ket vectors for modelling. As target of the mapping we take a real inner product space with dimension $n \leq k$. As extreme case we can map all values to the two-dimensional inner product space of a real one-qubit-system, see, for example, the mapping of three values in Fig. 2.

An intuitive question arises from where we get the right ket vectors. Starting point is a $k \times k$ similarity matrix $S = \{s_{ij}\}$ expressing the required gradual similarity values between all value pairs. For the construction of the ket vectors, the similarity matrix must meet the following properties:

- *Unit interval*: All values of the matrix are elements from [0, 1].
- *Diagonal values*: All diagonal values refer to the similarity of values to themselves and are therefore 1.
- *Symmetry*: The matrix is symmetric since similarity is usually required to be symmetric.
- *Square-rooted positive semi-definiteness*: For reasons explained in the sequel, we require the matrix of square roots $S^{\frac{1}{2}} := \{\sqrt{s_{ij}}\}$ to be positive semi-definite. That is, the eigenvalues must be non-negative.

Table 3 left shows an example of a similarity matrix.

Based on a similarity matrix S we can construct the ket vectors. First, we replace all matrix elements by their square roots yielding $S^{\frac{1}{2}}$. The motivation for this is that the projection probability given by quantum measurement corresponds to a squared inner product. Second, we perform a spectral decomposition of $S^{\frac{1}{2}}$ and obtain the matrix V containing orthonormal eigenvectors as rows and a diagonal matrix L with the corresponding non-negative eigenvalues:

$$S^{\frac{1}{2}} = V^{\dagger} \cdot L \cdot V.$$

Table 3 Similarity values (left) and their element-wise square roots (right)

S	V_1	V_2	V_3
V_1	1	0.5	0
V_2	0.5	1	0.5
V_3	0	0.5	1

$S^{\frac{1}{2}}$	V_1	V_2	V_3
V_1	1	$1/\sqrt{2}$	0
V_2	$1/\sqrt{2}$	1	$1/\sqrt{2}$
V_3	0	$1/\sqrt{2}$	1

Since L is a diagonal matrix with non-negative values we can write it as a product of its square roots $L = L^{\frac{1}{2}} \cdot L^{\frac{1}{2}}$ and obtain:

$$\begin{aligned} S^{\frac{1}{2}} &= V^{\dagger} \cdot L^{\frac{1}{2}} \cdot L^{\frac{1}{2}} \cdot V \\ &= V^{\dagger} \cdot L^{\frac{1}{2}\dagger} \cdot L^{\frac{1}{2}} \cdot V \\ &= \left(L^{\frac{1}{2}} \cdot V\right)^{\dagger} \cdot L^{\frac{1}{2}} \cdot V \\ &= K^{\dagger} \cdot K, \end{aligned}$$

with $K = \{k_{ij}\} = L^{\frac{1}{2}} \cdot V$. The columns of matrix K correspond to the required ket vectors. However, they are vectors of k dimensions. The number of dimensions is usually higher than necessary. Let us inspect the diagonal matrix L containing the eigenvalues. Very often, some of the eigenvalues are zero. The corresponding dimensions can therefore be removed and we end up with ket vectors of an inner product space of a dimension n less than k.[1] The mapping is given by:

$$\begin{aligned} QDom(\langle \mathtt{dt} \rangle) &= \{|\mathrm{V}_1\rangle, \ldots, |\mathrm{V}_k\rangle\} \\ Dom(\langle \mathtt{dt} \rangle) &\rightarrow QDom(\langle \mathtt{dt} \rangle) \\ \forall j \in [1, k] : \mathrm{V}_j \mapsto |\mathrm{V}_j\rangle &= \sum_{i=1}^{n} k_{ij} |i\rangle \in \mathrm{span}\{|1\rangle, \ldots, |n\rangle\} \doteq \mathbb{R}^n \end{aligned}$$

where $|i\rangle$ denotes the i-th canonical unit vector of $\mathbb{R}^n$.

We will demonstrate the derivation of ket vectors from a similarity matrix using the example given in Table 3. The similarity matrix is given left and its square root is given right. The Cholesky-decomposition yields the square matrix given in Table 4. The matrix can be reduced by the last row since the corresponding eigenvalue is zero. Thus, we obtain three two-dimensional ket vectors from the resulting columns. They are illustrated in Fig. 3.

[1] A more efficient method to derive the ket vectors is to apply the Cholesky-decomposition from $S^{\frac{1}{2}}$ [5].

Table 4 Ket vectors as columns

$\lvert V_1\rangle$	$\lvert V_2\rangle$	$\lvert V_3\rangle$
1	$1/\sqrt{2}$	0
0	$1/\sqrt{2}$	1
0	0	0

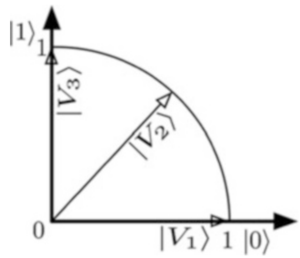


Fig. 3 Ket vectors as one-qubit-vectors

$$
\begin{aligned}
\lvert V_i\rangle &\ \dot{\in}\ \mathbb{R}^2 \\
\lvert V_1\rangle &\doteq \begin{pmatrix}1\\0\end{pmatrix} \\
\lvert V_2\rangle &\doteq \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \\
\lvert V_3\rangle &\doteq \begin{pmatrix}0\\1\end{pmatrix}.
\end{aligned}
$$

Take a ket vector $\lvert V_i\rangle$ expressing a non-orthogonal property value. We use a projector for a similarity measurement '$\approx$' of $\lvert V_i\rangle$ against $\lvert V_x\rangle$. The projector is defined as $P = \lvert V_i\rangle\langle V_i\rvert$. The measurement result corresponds to the squared cosine of the enclosed angle:

$$\langle V_x\rvert P\lvert V_x\rangle = \langle V_x\vert V_i\rangle\langle V_i\vert V_x\rangle.$$

In our example in Tables 1 and 2, the properties `fuel tank` and `kilometre` are non-orthogonal. Condition `FT1` can be evaluated by applying the projector $P = \lvert V_{35}\rangle\langle V_{35}\rvert$ against a ket vector encoding the fuel tank of a certain car.

4 Data Type Construction

Elementary data types are often not powerful enough to model complex data structures for many practical applications. Instead, there is a need to construct complex data structures on top of elementary data types.

In our example in Table 1 we have a complex data structure: the car dealership involves a set of cars, a service booklet contains a list of service entries, and properties of the engine are grouped together.

The type theory of object-oriented databases [6] and also most imperative programming languages exploit data type constructors for construction of complex data types based on existing data types given as constructor arguments. The two most important data type constructors are:

– `tuple`: The tuple data type constructor groups a fixed number of properties which we will call *components*. Every component is composed of a data type and an identifying label:

$$\langle \texttt{tuple-dt} \rangle := \texttt{tuple}\,(\text{name}_1 : \langle \texttt{Dt}_1 \rangle\,, \ldots, \text{name}_N : \langle \texttt{Dt}_N \rangle)\,.$$

The domain of the constructed data type is the result from multiplying (Cartesian) the domains of the given data types:

$$\mathit{Dom}\,(\langle \texttt{tuple-dt} \rangle) := \mathit{Dom}\,(\langle \texttt{Dt}_1 \rangle) \times \ldots \times \mathit{Dom}\,(\langle \texttt{Dt}_N \rangle)\,.$$

A tuple value needs to be constructed analogously to the data type construction. If $\texttt{value}_1$ to $\texttt{value}_N$ are values of the properties, then the tuple value is defined by:

$$\texttt{tuple-value} := \texttt{tuple}\,(\texttt{value}_1, \ldots, \texttt{value}_N)\,.$$

For a given tuple value we can access a certain property name_i using the dot-operator: $\texttt{tuple-value}\,.\,\text{name}_i$.

– `set`: The set data type constructor creates a data type for a set of values from the given data type:

$$\langle \texttt{set-dt} \rangle := \texttt{set}\,(\langle \texttt{dt} \rangle)\,.$$

The domain of the constructed data type is given by the power set:

$$\mathit{Dom}\,(\langle \texttt{set-dt} \rangle) := \mathcal{P}(\mathit{Dom}\,(\langle \texttt{dt} \rangle)).$$

The initial set value is constructed by:

$$\texttt{set-value} := \texttt{set}\,()\,.$$

This operation creates an empty set. Further operation for inserting and removing of elements and general set operations are available but not detailed here.

Using our example we demonstrate data type construction. Assume that the following elementary data types $\langle\texttt{lt-dt}\rangle$,[2] $\langle\texttt{yoc-dt}\rangle$,[3] $\langle\texttt{kilometre-dt}\rangle$, and $\langle\texttt{date-dt}\rangle$ are given. The data types for a car and for an entry of the service booklet are constructed as follows:

$$\langle\texttt{car-dt}\rangle := \texttt{tuple}\,(\text{lt} : \langle\texttt{lt-dt}\rangle, \text{yoc} : \langle\texttt{yoc-dt}\rangle)$$

$$\langle\texttt{entry-dt}\rangle := \texttt{tuple}\,(\text{kilometre} : \langle\texttt{kilometre-dt}\rangle, \text{date} : \langle\texttt{date-dt}\rangle)\,.$$

Data type construction is recursive. Starting points are elementary data types:

– An elementary data type is a *data type*.
– Let $\langle\texttt{dt}_1\rangle$ to $\langle\texttt{dt}_N\rangle$ be data types and name_1 to name_N labels. Then $\texttt{tuple}\,(\text{name}_1 : \langle\texttt{dt}_1\rangle\,, \ldots, \text{name}_N : \langle\texttt{dt}_N\rangle)$ is a *data type*.
– If $\langle\texttt{dt}\rangle$ is a data type, then $\texttt{set}\,(\langle\texttt{dt}\rangle)$ is a *data type*.

In our example we define the data type $\langle\texttt{service-booklet-dt}\rangle$ and the data type $\langle\texttt{car-dealership-dt}\rangle$ by applying the set data type constructor on top of the tuple data type constructor:

$$\langle\texttt{service-booklet-dt}\rangle := \texttt{set}\,(\langle\texttt{entry-dt}\rangle)$$

$$\langle\texttt{car-dealership-dt}\rangle := \texttt{set}\,(\langle\texttt{car-dt}\rangle)\,.$$

Analogously we can define all data types required for our example.

5 Quantum-Based Data Type Constructors

After introducing data type constructors we will map them onto concepts of quantum mechanics. The idea is to define a bijection that relates values of constructed data types to vectors of an inner product space.

5.1 *Tuple Data Type Constructor*

The tuple data type constructor groups a fixed number of components. Every component consists of a label and a data type. Let us be given the labels name^j and the values $V^j_{j_i} \in Dom(\langle\texttt{dt}^j\rangle)$ of the components j as well as the corresponding ket vectors $|\text{V}^j_{j_i}\rangle$. Following the mathematics of quantum mechanics, several inner

[2] $\langle\texttt{lt-dt}\rangle$ stands for $\langle\texttt{license tag}\rangle$.
[3] $\langle\texttt{yoc-dt}\rangle$ stands for $\langle\texttt{year of construction}\rangle$.

product spaces are combined to a larger one by applying the tensor product. For a tuple data type construction

$$\langle \texttt{T-dt} \rangle := \texttt{tuple}\left(\text{name}^1 : \langle \texttt{dt}^1 \rangle, \ldots, \text{name}^N : \langle \texttt{dt}^N \rangle\right),$$

the ket vectors and their inner product spaces are combined by use of the tensor product:

$$\begin{aligned} QDom(\langle \texttt{T-dt} \rangle) &= QDom(\langle \texttt{dt}^1 \rangle) \otimes \ldots \otimes QDom(\langle \texttt{dt}^N \rangle) \\ Dom(\langle \texttt{T-dt} \rangle) &\rightarrow QDom(\langle \texttt{T-dt} \rangle) \\ \left(\mathrm{V}^1_{1_i}, \ldots, \mathrm{V}^N_{N_i}\right) &\mapsto |\mathrm{V}^1_{1_i}\rangle \otimes \ldots \otimes |\mathrm{V}^N_{N_i}\rangle. \end{aligned}$$

The ket vector $|\mathrm{V}^1_{1_i}\rangle \otimes \ldots \otimes |\mathrm{V}^N_{N_i}\rangle$ can be shortly notated as $|\mathrm{V}^1_{1_i} \ldots \mathrm{V}^N_{N_i}\rangle$. If d^l is the number of basis ket vectors of component l, that is, the number of dimensions, then the tensor product needs $d^1 \cdot \ldots \cdot d^N$ many basis ket vectors. That is, with respect to the tensor product the number of dimensions is multiplied and not added as in the case of the Cartesian product.

Take a three-component tuple value $|\mathrm{V}^1\mathrm{V}^2\mathrm{V}^3\rangle$ as a result of a tuple type construction. Furthermore, let $P^1 = \sum_{i^1} |i^1\rangle\langle i^1|$, $P^2 = \sum_{i^2} |i^2\rangle\langle i^2|$, and $P^3 = \sum_{i^3} |i^3\rangle\langle i^3|$ be projectors of the respective components. For measuring the tuple value also the projectors need to be combined:

$$\begin{aligned} P^{123} &= P^1 \otimes P^2 \otimes P^3 \\ &= \left(\sum_{i^1} |i^1\rangle\langle i^1|\right) \otimes \left(\sum_{i^2} |i^2\rangle\langle i^2|\right) \otimes \left(\sum_{i^3} |i^3\rangle\langle i^3|\right) \\ &= \sum_{i^1}\sum_{i^2}\sum_{i^3} |i^1 i^2 i^3\rangle\langle i^1 i^2 i^3|. \end{aligned}$$

The measurement is performed as

$$\langle \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3 | P^{123} | \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3\rangle = \sum_{i^1}\sum_{i^2}\sum_{i^3} \langle \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3 | i^1 i^2 i^3\rangle\langle i^1 i^2 i^3 | \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3\rangle.$$

Due to

$$\langle \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3 | i^1 i^2 i^3\rangle = \langle \mathrm{V}^1 | i^1\rangle\langle \mathrm{V}^2 | i^2\rangle\langle \mathrm{V}^3 | i^3\rangle$$

we obtain

$$\langle \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3 | P^{123} | \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3\rangle = \langle \mathrm{V}^1 | P^1 | \mathrm{V}^1\rangle\langle \mathrm{V}^2 | P^2 | \mathrm{V}^2\rangle\langle \mathrm{V}^3 | P^3 | \mathrm{V}^3\rangle.$$

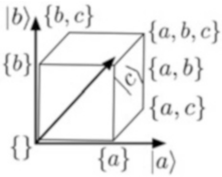


Fig. 4 Sets as one of 2^3 corners over the $\{a, b, c\}$-cube (for illustration purpose, the distances between origin and the corners are not all one)

Thus, measuring a tuple value equals the product of the component measurements.

Sometimes we want to measure only one component. In that case we use the identity $\mathbf{1} = \sum_{|i\rangle \in ONB} |i\rangle\langle i|$ as projector for the no-care-components where ONB stands for orthonormal basis:

$$\langle \mathrm{V}^1\mathrm{V}^2\mathrm{V}^3| \left(P^1 \otimes \mathbf{1} \otimes \mathbf{1}\right) |\mathrm{V}^1\mathrm{V}^2\mathrm{V}^3\rangle = \langle \mathrm{V}^1|P^1|\mathrm{V}^1\rangle.$$

5.2 Set Data Type Constructor

The set data type constructor is used on top of an orthonormal data type $\langle\mathtt{dt}\rangle$ with $QDom(\langle\mathtt{dt}\rangle) = \{|V_1\rangle, \ldots, |V_k\rangle\}$ being an orthonormal basis:

$$\langle\mathtt{S\text{-}dt}\rangle := \mathtt{set}(\langle\mathtt{dt}\rangle).$$

Every element of $Dom(\langle\mathtt{S\text{-}dt}\rangle)$ is a subset of $Dom(\langle\mathtt{dt}\rangle)$. The idea of our quantum modelling approach is the usage of the superposition principle, see Fig. 4. To construct the vector space $QDom\ (\langle\mathtt{S\text{-}dt}\rangle)$, we have to collect all superpositions of ket vectors from $QDom\ (\langle\mathtt{dt}\rangle)$. Mathematically, this leads to the vector space of all linear combinations of ket vectors from $QDom\ (\langle\mathtt{dt}\rangle)$, which is called the span of $QDom\ (\langle\mathtt{dt}\rangle)$. The ket vectors of the set elements are superimposed:

$$QDom\ (\langle\mathtt{S\text{-}dt}\rangle) = \mathrm{span}\,(QDom\ (\langle\mathtt{dt}\rangle)) \quad (1)$$

$$Dom\ (\langle\mathtt{S\text{-}dt}\rangle) \to QDom\ (\langle\mathtt{S\text{-}dt}\rangle) \quad (2)$$

$$Dom\ (\langle\mathtt{S\text{-}dt}\rangle) \ni S \mapsto \frac{1}{\sqrt{|S|}} \sum_{\mathrm{V}_i \in S} |\mathrm{V}_i\rangle = |S\rangle \in QDom\ (\langle\mathtt{S\text{-}dt}\rangle)\,. \quad (3)$$

The mapping does not change the number of basis ket vectors, that is, the number of dimensions.[4] Please note that the set data type constructor yields a non-orthogonal data type.

Let us be given a superposition ket vector $|S\rangle$. We want to test by quantum measurement if the value V_j mapped to $|\mathrm{V}_j\rangle \in QDom(\langle\mathtt{dt}\rangle)$ is a member of the set S. For measurement we use the projector $P = |\mathrm{V}_j\rangle\langle\mathrm{V}_j|$:

[4]The special case of an empty set is discussed later on.

Table 5 Measurement of set $S = \{V_1, V_2, V_3\}$ with the set $V = \{V_2, V_3, V_4, V_5\}$

	V_2	V_3	V_4	V_5
V_3	0	1/3	0	0
V_2	1/3	0	0	0
V_1	0	0	0	0

$$\langle S|V_j\rangle\langle V_j|S\rangle = \frac{1}{|S|}\sum_{V_i \in S}\langle V_i|V_j\rangle\langle V_j|V_i\rangle = \begin{cases} \frac{1}{|S|} & \text{if } V_j \in S \\ 0 & \text{otherwise.} \end{cases}$$

A measurement with a set-valued projector $P = \sum_{V_j \in V} |V_j\rangle\langle V_j|$ for the set $V \subseteq Dom(\langle \texttt{S-dt}\rangle)$ yields:

$$\langle S|P|S\rangle = \frac{1}{|S|}\sum_{V_i \in S}\sum_{V_j \in V}\langle V_i|V_j\rangle\langle V_j|V_i\rangle = \frac{|V \cap S|}{|S|}.$$

Table 5 illustrates an example for the set measurement. As a result we obtain $2/3 = 1/3 + 1/3$.

The empty set needs special consideration. The problem is that there is a superposition of nothing, that is, the empty set cannot be represented by any vector of the given data type. Therefore, we insert an additional basis vector $|NULL\rangle$ into $QDom\,(\langle \texttt{dt}\rangle)$. The ket vector $|NULL\rangle$ is only used if an empty set needs to be encoded.

Assume a non-orthogonal data type $\langle \texttt{no-dt}\rangle$ with

$$QDom(\langle \texttt{no-dt}\rangle) = \{|V_1\rangle, \ldots, |V_k\rangle\}$$

is given. However, the set data type constructor is not defined for non-orthogonal data types. One solution is to orthogonalize the non-orthogonal data type. This can be easily realized by applying a tuple construction together with an auxiliary orthogonal data type $\langle \texttt{aux-dt}\rangle$ of dimension k. So, you obtain the orthogonalized data type $\langle \texttt{o-dt}\rangle$ by constructing:

$$\langle \texttt{o-dt}\rangle := \texttt{tuple}\,(\text{id} : \langle \texttt{aux-dt}\rangle\,, \text{value} : \langle \texttt{no-dt}\rangle)$$

and assigning bijectively a unique number from $dom(\langle \texttt{aux-dt}\rangle)$ to every value from $dom(\langle \texttt{no-dt}\rangle)$:

$$QDom\,(\langle \texttt{o-dt}\rangle) \subseteq \{|1\rangle, \ldots, |k\rangle\} \otimes QDom\,(\langle \texttt{no-dt}\rangle)$$

$$QDom(\langle \texttt{no-dt}\rangle) \to QDom(\langle \texttt{o-dt}\rangle)$$

$$|V_i\rangle \mapsto |i\rangle \otimes |V_i\rangle.$$

6 Conclusion

A relational database is based on the concept of a *relation* [7]. A relation is defined as a set (`set`) of tuples (`tuple`) containing property values. A relational database itself can be seen as a set (`set`) of tuples (`tuple`) containing the name of a relation and the relation itself. Thus, the data structure of a relational database can be expressed by our recursively defined *data type*. Since a relational database is a universal data structure for modelling arbitrary real-world scenarios, our recursively defined concept of a data type is also universal.

In this work we focused on the two most important data type structures: the tuple and the set data type constructor. Please note that further data type constructors like list, bag, dictionary, and array can be easily simulated using the tuple and the set data type constructors, see relational database design [7].

An interesting question for further work is how to express integrity constraints in quantum mechanics. For example, how can we incorporate the concepts of uniqueness and functional dependencies into the world of quantum mechanics?

References

1. Nielsen, M. A., & Chuang, I. L. (2011). *Quantum computation and quantum information: 10th anniversary edition* (10th ed.). New York, NY: Cambridge University Press.
2. Schmitt, I., Römer, R., Wirsching, G., & Wolff, M. (2017). Denormalized quantum density operators for encoding semantic uncertainty in cognitive agents. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 000165–000170). Piscataway: IEEE.
3. Schmitt, I. (2008). QQL: A DB&IR query language. *The VLDB Journal, 17*, 39–56.
4. Birkhoff, G., & Von Neumann, J. (1936). The logic of quantum mechanics. *Annals of Mathematics, 37*, 823–843.
5. Higham, N. J. (2009). Cholesky factorization. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*, 251–254.
6. Bancilhon, F. (1997). Object-oriented databases. In A. B. Tucker (Ed.), *The computer science and engineering handbook* (pp. 1158–1170). Boca Raton, FL: CRC Press.
7. Elmasri, R., & Navathe, S. (2010). *Fundamentals of database systems*. Boston, MA: Addison-Wesley Publishing Company.

Incorporating Weights into a Quantum-Logic-Based Query Language

Ingo Schmitt

Abstract Traditional database query languages are based on set theory and crisp first order logic. However, many applications require imprecise conditions which return result objects associated with a degree of fulfillment. For example, a research paper should be reviewed by giving a degree of fulfillment for originality and relevance. Very often, imprecise conditions of a query are of different importance. Thus, a query language should allow the user to give conditions different weights. This paper proposes a weighting approach which is realized by means of conjunction, disjunction, and negation. Thus, our weighting is completely embedded within a logic. As a result, logical rules are preserved and can be used for query reformulation and optimization. As underlying logic-based query language, we use the CQQL query language. Furthermore we demonstrate that our weighting approach is applicable to further logic-based query languages.

Keywords Weights · Database query language · Information retrieval · DB&IR

1 Introduction

Evaluating a traditional database query against an object returns `true` on match and `false` on mismatch. Unfortunately, there are many application scenarios where such an evaluation is impossible or does not adequately meet user's needs. For example, a user may search for a very comfortable but inexpensive camcorder by means of a query containing these unrealistic requirements. Obviously, he wants to see how *close* certain product offers are to his vision. Another example is a text retrieval search where in general an exact match is impossible. Thus, there is a need for incorporating impreciseness and proximity into a logic-based database query

I. Schmitt (✉)
Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany
e-mail: schmitt@b-tu.de

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_7

language. Objects fulfill such a query to a certain degree expressed by a result value out of the interval [0, 1] that is used for an object ranking. Many proposed query languages for such scenarios use concepts from fuzzy logic [1]. Another approach is based on quantum logic [2].

Using more than one imprecise query conditions in a query often causes a new problem: which importance on the result should one condition have in comparison to another one? For example, for the reviewing process of research papers gradual fulfillment of criteria like originality, technical quality, and relevance must be combined into one overall score bringing all paper submissions into one sequence. The program committee decides on the impacts (weights) of the used criteria.

In the following, we develop a weighting scheme for database queries based on a logic that may involve Boolean database, proximity, and retrieval conditions. In other words, we discuss queries bridging the retrieval and the database world. The corresponding queries are intended to be formulated by database experts.

Cooper [3] discusses alternatives to the Boolean model for information retrieval and the need for a weighting. There, the concept of *weighted request terms* is very similar to our proposed weighting approach.

One well-known approach for incorporating weights into scoring rules was published by Fagin and Wimmers [4]. However, this approach and similar approaches introduce a *weighting formula* applied on top of an existing query language. Being outside the original query language leads to the following problem: The rules of the underlying logic are not valid anymore. For example, associativity for a Fagin-weighted query does not longer hold.

In this work we develop a weighting approach and define a sound weighting semantics. Our main contributions are:

1. We will show that a logic-based query language like CQQL [2] is powerful enough to support weights on conditions. We map weighted conjunction and weighted disjunction to their unweighted counterparts within the original query language. In contrast to other approaches, we are not restricted to the fuzzy t-norm/t-conorm `min/max` [1].
2. Since our approach is completely embedded within a logic existing logical rules remain and can be used for further query transformations.
3. Our approach allows the formulation of *connected weights*. Surprisingly, its evaluation returns a weighted sum although only conjunction, disjunction, and negation are used.
4. We demonstrate that our weighting approach can be applied to a logic-based query language even if only Boolean conditions are involved.

The next section gives a motivating example. Section 3 states some basic definitions and requirements for a weighting approach. After briefly introducing the core of the language CQQL in Sect. 4, we present our approach in Sect. 5. Section 6 compares our approach with alternative approaches. The last section ends with a final conclusion.

2 A Motivating Example

This section introduces a motivating example which helps to understand our weighting problem. Assume, you are planning your next Summer holiday. The goal is to find a Summer cottage located on the shores of the Baltic Sea. Best time would be the middle of August. Furthermore, you look for a family-friendly cottage with a high star rating. Of course, it should not be too expensive.

Let the Summer cottage offers be encoded in XML. Available tags together with corresponding atomic conditions of our search query are listed in Table 1.

We distinguish three types of atomic conditions:

1. *text retrieval*: A text retrieval condition determines whether a given text is about a certain topic. The returned score value from the interval [0, 1] can be regarded as the estimated probability that the text is about the required topic. It is obtained by using techniques of information retrieval [5, 6], e.g., computing the cosine of the enclosed angle between two document vectors. We need this kind of condition since the information about family-friendliness is encoded as free text.
2. *database*: A database condition returns a Boolean value in dependence of an attribute value. Either the condition is *completely* fulfilled or not. In our example we test `area` on equality with the string "Baltic Sea."
3. *proximity*: A proximity condition returns a score value from [0, 1] in dependence how *strong* the attribute value fulfills the condition. The value 1 corresponds to an exact match, the value 0 to a maximal mismatch, and any value in between to a *gradual* fulfillment. In our example, we use a proximity condition to find out how near the star rating to the maximum rating is.

The atomic conditions in Table 1 are combined by a logic yielding a complex query. It combines all three kinds of atomic conditions. However, all conditions are then implicitly equally weighted. Instead, we may want to see the most comfortable Summer cottage where the price is of less importance than the star rating, or a Summer cottage which fits best the date condition. Figure 1 (left) lists some example offers with their score values from the corresponding conditions.[1] The first four cottages are good offers which should be displayed in dependence of certain weights.

Table 1 Atomic conditions

Name	Tag	Atomic condition	Type
T	Text	About "family-friendly"	Text retrieval
A	Area	= "Baltic Sea"	Database
D	Date	≈ "mid of August"	Proximity
S	Star rating	≈ max_rating	Proximity
P	Price	≈ max_price	Proximity

[1]The Boolean `true` is encoded by a 1 and `false` by a 0.

name	T	A	D	S	P
deluxe	1	1	0.6	1	0.8
cheap	0.6	1	0.6	0.4	0.2
fair	0.8	1	0.8	0.8	0.4
exact_date	0.8	1	1	0.8	0.5
too_expensive	0.8	1	0.8	0.8	0.8
bad_quality	0.6	1	0.4	0.2	0.2
not_baltic	1	0	0.8	0.6	0.6

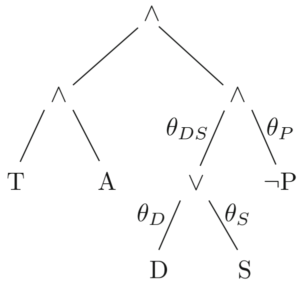


Fig. 1 Summer cottages and weighted condition tree

Figure 1 (right) depicts the chosen weighted query. The weight θ_D defines the importance of the date condition within the disjunction. Weights can also be assigned to non-atomic conditions: θ_{DS} gives a weight to a disjunctive branch of a conjunction. When we give `P` in contrast to `DS` a small weight our expectation is to obtain the cottage "deluxe." By using other weights the query system should find the cottages "cheap," "fair," and "exact_date," correspondingly.

We are faced with finding a formalism which consistently incorporates weights into a logic-based query system. The meaning of *consistency* is explained in more detail in the following section.

3 Basic Definitions and Requirements

In our example, we combine database, proximity, and text retrieval conditions. The degree of their fulfillment is encoded by a score value from the interval [0, 1]. This kind of encoding conforms with the theory of fuzzy logic [1], i.e., every atomic condition of our example over a database can be interpreted as a fuzzy set. In order to combine scores to a complex condition we need a score function. Fuzzy logic defines a *t-norm* for a conjunction and a *t-conorm* for a disjunction.

Definition 1 A function $\top : [0, 1]^2 \rightarrow [0, 1]$ is a *t-norm* if it satisfies the following rules:

$$\begin{array}{ll} \forall a \in [0,1] : \top(a,1) = a & \text{(identity element)} \\ \forall a,b,c \in [0,1] : a \leq b \Longrightarrow \top(a,c) \leq \top(b,c) & \text{(monotonicity)} \\ \forall a,b,c \in [0,1] : \top(a,\top(b,c)) = \top(\top(a,b),c) & \text{(associativity)} \\ \forall a,b \in [0,1] : \top(a,b) = \top(b,a) & \text{(commutativity).} \end{array}$$

Definition 2 A function $\bot : [0, 1]^2 \rightarrow [0, 1]$ is a *t-conorm* if $\bot$ is commutative, associative, monotone and 0 is its identity element.

Let us now define a *score function* for a logic expression.

Definition 3 Let $o \in O$ be a database object, $sv_{ac}(o) \in [0, 1]$ a score value obtained from evaluating an atomic condition ac on o, and φ a logic formula constructed by recursively applying $\wedge$, $\vee$, and $\neg$ on atomic conditions. A *score function* $f_\varphi : O \rightarrow [0, 1]$ based on $\top$ and $\bot$ is defined by

$$f_\varphi(o) = \begin{cases} sv_{ac}(O) & \text{if } \varphi = ac \\ \top\left(f_{\varphi_1}(o), f_{\varphi_2}(o)\right) & \text{if } \varphi = \varphi_1 \wedge \varphi_2 \\ \bot\left(f_{\varphi_1}(o), f_{\varphi_2}(o)\right) & \text{if } \varphi = \varphi_1 \vee \varphi_2 \\ 1 - f_{\varphi_1}(o) & \text{if } \varphi = \neg\varphi_1. \end{cases}$$

With respect to a logic formula φ a score function assigns to every object a value from [0, 1]. The result of a query over a logic formula φ is a list of all objects sorted descendingly by their score values.

Next we define a *weighted* score function.

Definition 4 Let $o \in O$ be a database object, $sv_{ac}(o) \in [0, 1]$ a score value obtained from evaluating an atomic condition ac on o, $\Theta = \{\theta_1, \ldots, \theta_n\}$ with $\theta_i \in [0, 1]$ a set of weights, and φ a logic formula constructed by recursively applying $\wedge$, $\vee$, $\wedge_{\theta_i,\theta_j}$, $\vee_{\theta_i,\theta_j}$, and $\neg$ on atomic conditions. A *weighted score function* $f_\varphi^\Theta : O \rightarrow [0, 1]$ based on $\top$, $\bot$, and the weight functions $w_{\theta_i,\theta_j}^\wedge$ and $w_{\theta_i,\theta_j}^\vee$ is defined by

$$f_\varphi^\Theta(o) = \begin{cases} sv_{ac}(o) & \text{if } \varphi = ac \\ \top\left(f_{\varphi_1}^\Theta(o), f_{\varphi_2}^\Theta(o)\right) & \text{if } \varphi = \varphi_1 \wedge \varphi_2 \\ \bot\left(f_{\varphi_1}^\Theta(o), f_{\varphi_2}^\Theta(o)\right) & \text{if } \varphi = \varphi_1 \vee \varphi_2 \\ w_{\theta_i,\theta_j}^\top\left(f_{\varphi_1}^\Theta(o), f_{\varphi_2}^\Theta(o)\right) & \text{if } \varphi = \varphi_1 \wedge_{\theta_i,\theta_j} \varphi_2 \\ w_{\theta_i,\theta_j}^\bot\left(f_{\varphi_1}^\Theta(o), f_{\varphi_2}^\Theta(o)\right) & \text{if } \varphi = \varphi_1 \vee_{\theta_i,\theta_j} \varphi_2 \\ 1 - f_{\varphi_1}^\Theta(o) & \text{if } \varphi = \neg\varphi_1. \end{cases}$$

Similar to [4] we state for the weight functions $w_{\theta_1,\theta_2}^\top$ and $w_{\theta_1,\theta_2}^\bot$ the following requirements:

- (R1) $\theta = 0$: If the weight of an operand is zero, then the operand has no impact on the result: $w_{0,1}^\top(a, b) = b$ (analogously for $\bot$ and the second operand). In our example, setting $\theta_P = 0$ means to ignore completely the price in order to obtain the cottage "deluxe."
- (R2) $\theta = 1$: A weight of one on both operands equals the unweighted case: $w_{1,1}^\top(a, b) = \top(a, b)$ (analogously for $\bot$). Setting all weights to 1 means the unweighted case where all conditions are equally weighted. In our example, we expect to obtain the cottage "fair."
- (R3) *continuity*: The weight functions are continuous on the weights: $\forall\theta_1', \theta_2' \in [0, 1] : \lim_{\theta_1 \rightarrow \theta_1'} \lim_{\theta_2 \rightarrow \theta_2'} w_{\theta_1,\theta_2}^\top(a, b) = w_{\theta_1',\theta_2'}^\top(a, b)$ (analogously for $\bot$). Small modifications of the price weight have small effects on the computed overall scores. There are no jumps.

– (R4) *embedding in a logic*: The weight function $w^{\top}_{\theta_1,\theta_2}$ obeys the same rules as the unweighted norms $\top$ (at least identity element, monotonicity, associativity, and commutativity):

$$\forall a, \theta \in [0,1] : w^{\top}_{1,\theta}(a, 1) = a;$$

$$\forall a, b, c, \theta, \theta_a, \theta_b, \theta_c \in [0,1] : w^{\top}_{\theta_a,\theta}(a, 1) \leq w^{\top}_{\theta_b,\theta}(b, 1) \implies w^{\top}_{\theta_a,\theta_c}(a, c) \leq w^{\top}_{\theta_b,\theta_c}(b, c);$$

$$\forall a, b, c, \theta_a, \theta_b, \theta_c, \theta_{bc} \in [0,1] : w^{\top}_{\theta_a,\theta_{bc}}\left(a, w^{\top}_{\theta_b,\theta_c}(b, c)\right) = w^{\top}_{1,\theta_{bc}\theta_c}(w^{\top}_{\theta_a,\theta_{bc}\theta_b}(a, b), c) \text{ and}$$

$$\forall a, b, \theta_a, \theta_b \in [0,1] : w^{\top}_{\theta_a,\theta_b}(a, b) = w^{\top}_{\theta_b,\theta_a}(b, a).$$

Analogously, the weight function $w^{\perp}_{\theta_1,\theta_2}$ must obey the rules, too. The logic rules provide us semantical equivalences of syntactically different conditions. In this way, we obtain a kind of independence from how (syntactically) a query is formulated by a user. Furthermore, such equivalences enable an optimization for query processing.

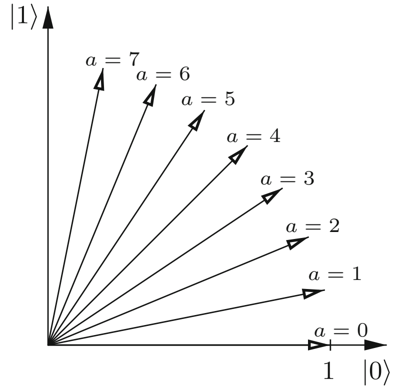
– (R5) *linearity:* The weight functions are linear on the weights: $w^{\top}_{\alpha\theta_1+(1-\alpha)\theta'_1,\theta_2}(a, b) = \alpha w^{\top}_{\theta_1,\theta_2}(a, b) + (1-\alpha) w^{\top}_{\theta'_1,\theta_2}(a, b)$ (analogously for $\perp$ and second weight). In our example, if we choose $\theta_P = 1/2$, then the condition $\overline{P}$ has exactly half of impact.

4 Conditions of the Logic-Based Similarity Query Language CQQL

In this section we sketch the construction and evaluation of conditions of the similarity query language CQQL [2]. CQQL stands for *commuting quantum query language* and is based on the mathematics behind quantum mechanics. This language enables the logic-based construction of queries from traditional database and information retrieval (including proximity) conditions. The underlying idea is to use the theory of vector spaces (linear algebra) for query processing. All attribute values of an object are embodied by the direction of a normalized object vector. The query itself corresponds to a vector subspace spanned by a set of orthonormal vectors. If the object vector belongs to the subspace, then we interpret it as a complete match. Otherwise we compute the minimal angle between them. Its squared cosine (squared scalar product) is interpreted as evaluation result. In [2] we show that this formalism covers three different types of conditions:

– *information retrieval*: A well-known retrieval model [6] is the vector space model which conforms with our formalism [7].

Fig. 2 Equiangular mapping



– *database condition*: A traditional database condition returns just Boolean values. This can be simulated by mapping bijectively every possible attribute value to exactly one basis vector. These vectors are mutually orthonormal. Thus, querying values produces only two possible results, a 1 on match and a 0 on mismatch.
– *proximity condition*: The values for a proximity condition are encoded by vectors of different non-orthogonal angles. For example, Fig. 2 depicts the encoding of eight values in a qubit. Following that encoding [2], the evaluation of a condition that requires an eight-value attribute c having the value b returns $\cos^2 |a-c|\pi/16$.

The starting point of CQQL conditions is a *set of commutative conditions*.

Definition 5 Let $A = \{a_j\}$ be a finite set of attributes and $AC = \{ac_i(a_j)\}$ be a finite set of atomic attribute conditions of the form "$a_j = value$" or proximity conditions of the form "$a_j \approx value$."[2] Let furthermore be given a function $vs(ac_i(a_j)) = \{\mathbf{ac_i^1}, \ldots, \mathbf{ac_i^{k_i}}\}$ which assigns to every condition a set of orthonormal vectors forming a vector subspace. The condition set AC is called *commutative* if $\forall ac_{i_1}(a_{j_1}), ac_{i_2}(a_{j_2}) \in AC : \big(type(ac_{i_1}(a_{j_1})) \neq d \wedge type(ac_{i_2}(a_{j_2})) \neq d \wedge j_1 = j_2\big) \Longrightarrow i_1 = i_2$ holds where $type$ returns d for a database condition and p or r for a proximity or retrieval condition, respectively.

Commutativity means that no two proximity or retrieval conditions "$a_j \approx value_1$" and "$a_j \approx value_2$" with $value_1 \neq value_2$ on the same attribute a_j are allowed.

Lemma 1 *Let AC be a commutative set of atomic conditions over $A = \{a_j\}$. The set $CVS(AC) = \bigcup_{ac_i(a_j) \in AC} vs(ac_i(a_j))$ is a set of mutually orthonormal vectors.*

The lemma is a direct consequence of how the mapping function vs is realized (for more details, see [2]). It is very essential because it means that every subset of $CVS(AC)$ spans a vector subspace and corresponds bijectively to a CQQL condition.

[2]Without loss of generality and for simplicity we ignore here atomic conditions on more than one attribute, which are discussed in [2].

Definition 6 Let AC be a commutative set of atomic conditions on $A = \{a_j\}$. Then a CQQL condition φ is recursively defined by

- $\varphi \stackrel{def}{=} ac_i(a_j) \in AC$
- $\varphi \stackrel{def}{=} (\varphi_1 \wedge \varphi_2)$
- $\varphi \stackrel{def}{=} (\varphi_1 \vee \varphi_2)$
- $\varphi \stackrel{def}{=} (\neg\varphi_1)$,

where φ_1, φ_2 are CQQL conditions.

Theorem 1 *All CQQL conditions over a commutative set of atomic conditions together with conjunction, disjunction, and negation form a Boolean algebra.*

Proof The function vs maps bijectively every condition to a subset of $CVS(CS)$. Conjunction, disjunction, and negation are mapped to their corresponding set operations:

$$\begin{aligned} vs(ac_i(a_j)) &\subseteq CVS(AC) \\ vs(\varphi_1 \wedge \varphi_2) = vs(\varphi_1) \cap vs(\varphi_2) &\subseteq CVS(AC) \\ vs(\varphi_1 \vee \varphi_2) = vs(\varphi_1) \cup vs(\varphi_2) &\subseteq CVS(AC) \\ vs(\neg\varphi_1) = CVS(AC) \setminus vs(\varphi_1) &\subseteq CVS(AC). \end{aligned}$$

A set together with these standard set operations and the orthocomplement for the negation is a Boolean algebra. □

The CQQL conjunction is also a valid t-norm and the disjunction a valid t-conorm. In contrast to known fuzzy norms, they fulfill all Boolean algebra rules including distributivity, absorption, and idempotence. An interesting result from [2] is the fact that a complex CQQL condition in a specific syntactical form can be evaluated by means of simple arithmetics.

Definition 7 Let the function att return for every condition the set of its restricted attributes. Following [2] we define for an object o the evaluation f of a conjunction and disjunction of non-overlapping conditions as well of an exclusive disjunction. The negation is just the subtraction from 1.

$$\begin{aligned} f_{\varphi_1 \wedge \varphi_2}(o) &= \top^{CQQL}(f_{\varphi_1}(o), f_{\varphi_2}(o)) = f_{\varphi_1}(o) \cdot f_{\varphi_2}(o) \text{ if } att(\varphi_1) \cap att(\varphi_2) = \emptyset. \\ f_{\varphi_1 \vee \varphi_2}(o) &= \bot^{CQQL}(f_{\varphi_1}(o), f_{\varphi_2}(o)) \\ &= f_{\varphi_1}(o) + f_{\varphi_2}(o) - f_{\varphi_1}(o) \cdot f_{\varphi_2}(o) \text{ if } att(\varphi_1) \cap att(\varphi_2) = \emptyset. \\ f_{\varphi_1 \vee \varphi_2}(o) &= \bot^{CQQL}(f_{\varphi_1}(o), f_{\varphi_2}(o)) \\ &= f_{\varphi_1}(o) + f_{\varphi_2}(o) \text{ if } \exists ac : \varphi_1 = \text{`}ac \wedge \phi_1\text{`} \wedge \varphi_2 = \text{`}(\neg ac) \wedge \phi_2\text{`} \\ f_{\neg\varphi}(o) &= 1 - f_{\varphi}(o). \end{aligned}$$

Since our language obeys the rules of a Boolean algebra we can transform every possible CQQL condition into the required syntactical form, e.g., the disjunctive normal form or the one-occurrence-form [8]. Schmitt [2] gives an algorithm performing this transformation.

Our example condition without weights is a CQQL condition being already in the required form. Following Definition 7 we obtain $sv_T(o) \cdot sv_A(o) \cdot (sv_D(o) + sv_S(o) - sv_D(o) \cdot sv_S(o))(1 - sv_P(o))$ for an object o to be evaluated.

5 Logic-Based Weighting

Our approach of weighting CQQL conditions is surprisingly simple. The idea is to transform a weighted conjunction or a weighted disjunction into a logical formula without weights:

$$w^{\wedge}_{\theta_1,\theta_2}(\varphi_1,\varphi_2) \Longrightarrow (\varphi_1 \vee \neg\theta_1) \wedge (\varphi_2 \vee \neg\theta_2) \quad \text{and} \quad w^{\vee}_{\theta_1,\theta_2}(\varphi_1,\varphi_2) \Longrightarrow (\varphi_1 \wedge \theta_1) \vee (\varphi_2 \wedge \theta_2).$$

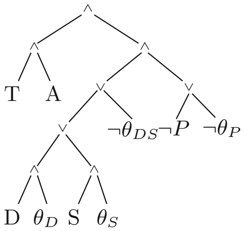
Definition 8 Let o be a database object, $sv_{ac}(o) \in [0, 1]$ a score value obtained from evaluating an atomic CQQL condition ac on o, $\Theta = \{\theta_1, \ldots, \theta_n\}$ with $\theta_i \in [0, 1]$ a set of weights, and φ a CQQL condition constructed by recursively applying $\wedge$, $\vee$, $\wedge_{\theta_i,\theta_j}$, $\vee_{\theta_i,\theta_j}$, and $\neg$ on a commuting set of atomic conditions. The *weighted score function* is defined by

$$f^{\Theta}_{\varphi}(o) = \begin{cases} f^{\Theta}_{(\varphi_1 \vee \neg\theta_i) \wedge (\varphi_2 \vee \neg\theta_j)}(o) & \text{if } \varphi = \varphi_1 \wedge_{\theta_i,\theta_j} \varphi_2 \\ f^{\Theta}_{(\varphi_1 \wedge \theta_i) \vee (\varphi_2 \wedge \theta_j)}(o) & \text{if } \varphi = \varphi_1 \vee_{\theta_i,\theta_j} \varphi_2 \\ \top^{CQQL}\left(f^{\Theta}_{\varphi_1}(o), f^{\Theta}_{\varphi_2}(o)\right) & \text{if } \varphi = \varphi_1 \wedge \varphi_2 \\ \bot^{CQQL}\left(f^{\Theta}_{\varphi_1}(o), f^{\Theta}_{\varphi_2}(o)\right) & \text{if } \varphi = \varphi_1 \vee \varphi_2 \\ 1 - f^{\Theta}_{\varphi_1}(o) & \text{if } \varphi = \neg\varphi_1 \\ sv_{ac}(o) & \text{if } \varphi = ac \\ sv_{ac_\theta} & \text{if } \varphi = \theta. \end{cases}$$

sv_{ac_θ} can be regarded as a special atomic condition without argument returning the constant θ.

Our weighting does not require a special weighting formula outside of the context of logic. Instead it uses exclusively the power of the underlying logic. Please notice that we can weight not only atomic conditions but also complex logical expressions. Thus, our approach supports a nested weighting.

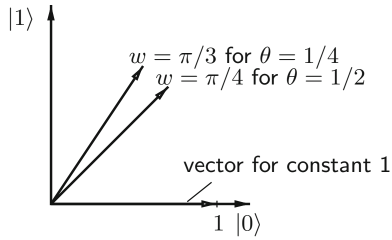
Figure 3 (left) shows the result of applying our mapping rules onto the weighted tree from Fig. 1. Following our CQQL evaluation we obtain the formula



name	θ_D	θ_S	θ_{DS}	θ_P	winner
equi_weight	1	1	1	1	fair
nonrelevant_price	1	1	1	0.1	deluxe
very_relevant_price	1	1	0	1	cheap
very_relevant_date	1	0.4	1	0.2	exact_date
zero_weight	0	0	0	0	cheap

Fig. 3 Expanded weighted condition tree and weight settings

Fig. 4 Weight mapping for $\theta = 1/4$ and $\theta = 1/2$



$$f_\varphi^\Theta(o) = sv_T(o) \cdot sv_A(o) \cdot sv_{DS\theta}(o) \cdot sv_{\neg P\theta_P}(o)$$

$$sv_{DS\theta}(o) = sv_{DS}(o) + (1 - sv_{\theta_{DS}}) - sv_{DS}(o) \cdot (1 - sv_{\theta_{DS}})$$

$$sv_{DS}(o) = sv_D(o) \cdot sv_{\theta_D} + sv_S(o) \cdot sv_{\theta_S} - sv_D(o) \cdot sv_{\theta_D} \cdot sv_S(o) \cdot sv_{\theta_S}$$

$$sv_{\neg P\theta_P}(o) = (1 - sv_P(o)) + (1 - sv_{\theta_P}) - (1 - sv_P(o)) \cdot (1 - sv_{\theta_P}).$$

The table in Fig. 3 (right) shows the chosen weight values and the corresponding cottages with the highest score value. Of course, for an end user it is not easy to find the right weight values. Instead, we propose the usage of linguistic variables like `very important`, `important`, `neutral`, `less important`, and `not important` and map them to numerical weight values. Another possibility is to use graphical weight sliders to adjust the importance of a condition.

Next, we show that our weighting approach fulfills our requirements.

Theorem 2 *The weight functions as defined in Definition 8 fulfill the requirements R1 to R5.*

Proof Following Theorem 1, CQQL conditions with conjunction, disjunction, and negations form a Boolean algebra. Our approach uses special atomic conditions ac_θ for different weights θ (Fig. 4).

As demonstrated, our weighting is realized inside the CQQL logic. Thus, it is easy to show the fulfillment of the requirements:

1. (R1) *weight=0*: $w_{0,1}^{\top}(a, b) = (a \vee \neg 0) \wedge (b \vee \neg 1) = 1 \wedge b = b$. The commutative variant and the disjunction are analogously fulfilled.

2. (R2) *weight=1*: $w^{\top}_{1,1}(a, b) = (a \vee \neg 1) \wedge (b \vee \neg 1) = a \wedge b = \top^{CQQL}(a, b)$ (analogously for $\bot$).
3. (R3) *continuity*: The weight functions are continuous on the weights since the underlying squared cosine function, conjunction, and disjunction are continuous.
4. (R4) *embedding in a logic*: Since our weighting is realized inside the logic formalism the resulting logic is still a Boolean algebra. The proof of the four formulas is thus straightforward.
5. (R5) *linearity*: The evaluation of a weighted CQQL conjunction and disjunction w.r.t. one weight is based on linear formulas.

So far, we applied our weighting approach to retrieval and proximity conditions of our CQQL language. What about applying it exclusively to traditional database conditions returning just Boolean values? In that case, we distinguish between two cases:

1. *Non-Boolean weights*: If the weights come from the interval [0, 1], then we cannot use the normal Boolean disjunction and conjunction. Instead we propose to replace them with the algebraic sum $(x + y - x \cdot y)$ and the algebraic product $(x \cdot y)$. This case can be regarded as a special case (only database conditions) of CQQL and it fulfills the requirements (R1) to (R5) if each condition is not weighted more than once. Otherwise we transform, see [2], that formula into the required form. Table 2 shows the effect of non-Boolean weights on database conditions whose evaluations are expressed by x and y.
2. *Boolean weights*: If the weights are Boolean values, then we can evaluate a weighted formula by using Boolean conjunction and disjunction fulfilling requirements (R1), (R2), and (R4). Requiring continuity (R3) and linearity (R5) on Boolean weights is meaningless. Boolean weights have the effect of a switch. A weighted condition can be switched to be active or inactive.

At the end, we present $(x \wedge \theta) \vee (y \wedge \neg\theta)$ as an interesting special case where one *connected* weight instead of two weights is used. In this case, the resulting evaluation formula is the weighted sum: $\theta * sv_x(o) + (1 - \theta) * sv_y(o)$. Very surprisingly, next logical transformation[3] shows that connected weights of a conjunction equal exactly connected weights of a disjunction.

Table 2 Weights on Boolean conditions

x	**y**	$x \wedge_{\theta,1} y := \left(x + \overline{\theta} - x\overline{\theta}\right) y$	$x \vee_{\theta,1} y := \theta x + y - \theta xy$
0 (false)	0 (false)	0 (false)	0 (false)
0 (false)	1 (true)	$\overline{\theta}$	1 (true)
1 (true)	0 (false)	0 (false)	θ
1 (true)	1 (true)	1 (true)	1 (true)

[3] For a simple notation, conjunction is expressed by a multiplication and disjunction by an addition symbol.

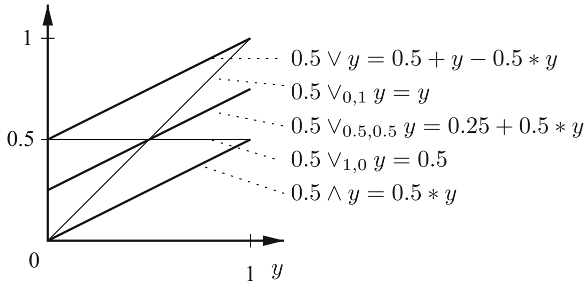


Fig. 5 Connected weights between conjunction and disjunction

$$(x + \neg\theta)(y + \theta) = xy + x\theta + y\neg\theta + \neg\theta\theta = xy\theta + xy\neg\theta + x\theta + y\neg\theta$$
$$= (xy + x)\theta + (xy + y)\neg\theta = x\theta + y\neg\theta.$$

This effect can be interpreted as a neutral combination of conditions. Figure 5 illustrates for a constant $x = 0.5$ that $x \vee_{\theta,1-\theta} y$ lies exactly in the middle between conjunction and disjunction.

6 Related Work

Weighting of non-Boolean query conditions is a well-known problem. Following [9], we distinguish four types of weight semantics: (1) weight as measure of importance of conditions, (2) weight as a limit on the amount of objects to be returned, (3) weight as threshold value, cf. [10, 11], and (4) weight as specification of an ideal database object. Next, we discuss related papers about weights as measure of importance. All logic-based weighting approaches supporting vagueness use the fuzzy t-norm/t-conorm `min/max`.[4] In contrast, our approach is a general logic-based approach where `min/max` is just one special case.

Fagin's Approach Fagin and Wimmers [4] propose a special arithmetic weighting formula to be applied on a score rule S. Let $\Theta = \{\theta_1, \dots, \theta_n\}$ be weights with $\theta_i \in [0, 1]$, $\sum_i \theta_i = 1$, and $\theta_1 \geq \theta_2 \geq \dots \geq \theta_n$. The weighted version of the score is then defined as $S^{\Theta}(\mu_1(o), \dots, \mu_n(o), \theta_1, \dots, \theta_n) = (\theta_1 - \theta_2)S(\mu_1(o)) + 2*(\theta_2 - \theta_3)$ $S(\mu_1(o), \mu_2(o)) + \dots + n*\theta_n S(\mu_1(o), \dots, \mu_n(o))$. Since this weighting approach is on top of a logic it violates R4. Thus, associativity in combination with `min/max`, for example, cannot be guaranteed (see [12]). Furthermore, Sung [13] describes a so-called stability problem for Fagin's approach. We can show that Fagin's formula with weights θ_1^F, θ_2^F applied to our CQQL logic can be simulated by our weighting where $\theta_1 = 1$ and $\theta_2 = 2 * \theta_2^F$ hold.

[4] `min/max` is the only fuzzy t-norm/t-conorm which supports idempotence and distributivity.

Arithmetic Formula on Operands The main idea of [14–16] is to apply arithmetical formulas on operands of a disjunction or conjunction. Thus, they all do not fulfill R4. For example, Sung [13] defines $S^{\Theta}(\mu_1(o), \ldots, \mu_n(o), \theta_1, \ldots, \theta_n) = S(\theta_1\mu_1(o), \ldots, \theta_n\mu_n(o))$. Interestingly, the operand formula $1-\theta(1-\mu(o))$ for the t-norm `min` proposed by Carson et al. [16] is very near to our evaluation formula on CQQL.

Weighted Sum Singitham et al. [17], Fuhr and Großjohann [18], and Oracle Corporation [19] propose a weighted sum approach completely independent from a logic. As shown, we can simulate the weighted sum by connected weights.

Logic-Based Weighting Approach on `min/max` The approaches proposed by Dubois and Prade [20], Pasi [21], and Yager [22] are very similar to our weighting approach and fulfill R1, R2, R3, and R4. However, they are strictly connected to the t-norm/t-conorm `min/max`. This results in a problem described by Fagin and Wimmers [4]. First, linearity cannot be fulfilled. Second, if $\mu_1(o) \geq 1 - \theta_2/\theta_1 \geq \mu_2(o)$ holds, then the result is completely independent from $\mu_1(o)$ and $\mu_2(o)$.

OWA Approach The OWA approach is discussed, for example, in [9, 23]. It was not developed to weight certain conditions. Instead, the user can assign weights to the highest score, to the second highest score, and so on. As a result, the characteristic of a conjunction can be gradually shifted to one of a disjunction. Thus, the weight does not express the importance of a condition.

We conclude that our weighting approach can be seen as a logic-based generalization of existing weighting approaches.

7 Conclusion

In this paper we propose a logic-based weighting mechanism which is completely embedded in a logic. As logic we used the logic of our query language CQQL. Later, we show that our approach can also be applied to other logics. A very interesting result are connected weights in CQQL. They produce the weighted sum by means of logic.

One problem not tackled here is the question, if a user is always able to specify weight values. We propose to use a kind of user interactions to infer that values. For example, a user starts with equal weight values and is able to adjust them after she/he sees the query result. Another approach is to learn weight values from required preferences over result objects.

We evaluated our approach successfully in a content-based image retrieval context [24]. There, weight values are not given by users but are learnt from user interactions.

References

1. Zadeh, L. A. (1988). Fuzzy logic. *IEEE Computer, 21*, 83–93.
2. Schmitt, I. (2008). QQL: A DB&IT query language. *The VLDB Journal: The International Journal on Very Large Data Bases, 17*, 39–56.
3. Cooper, W. S. (1988). Getting beyond Boole. *Information Processing & Management, 24*, 243–248.
4. Fagin, R., & Wimmers, E. L. (2000). A formula for incorporating weights into scoring rules. *Theoretical Computer Science, 239*, 309–338.
5. van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.
6. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Essex : ACM Press.
7. van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge: Cambridge University Press.
8. Olteanu, D., Huang, J., & Koch, C. (2009). Sprout: Lazy vs. eager query plans for tuple-independent probabilistic databases. In *2009 IEEE 25th International Conference on Data Engineering* (pp. 640–651).
9. Herrera-Viedma, E., Lopez-Herrera, A. G., Alonso, S., Porcel, C., & Cabrerizo, F. J. (2007). A linguistic multi-level weighted query language to represent user information needs. In *Proceedings of the IEEE International Conference on Fuzzy Systems, London, July 23–26, 2007* (pp. 1–6). Piscataway: IEEE.
10. Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 129–146.
11. Van Rijsbergen, C. J., Robertson, S. E., & Porter, M. F. (1980). *New models in probabilistic information retrieval*. London: British Library Research and Development Department London.
12. Schulz, N., & Schmitt, I. (2003). Relevanzwichtung in komplexen Ähnlichkeitsanfragen. In G. Weikum, H. Schöning, & E. Rahm (Eds.), *Datenbanksysteme in Business, Technologie und Web, BTW'03, 10. GI-Fachtagung, Leipzig, Februar 2003. Lecture notes in informatics (LNI)* (Vol. P-26, pp. 187–196). Bonn: Gesellschaft für Informatik.
13. Sung, S. (1998). *A Linear Transform Scheme for Combining Weights into Scores*. Technical report, Rice University.
14. Bookstein, A. (1980). Fuzzy requests: An approach to weighted Boolean searches. *Journal of the American Society for Information Science (JASIS), 31*, 240–247.
15. Bookstein, A. (1980). A comparison of two weighting schemes for Boolean retrieval. In R. N. Oddy, S. E. Robertson, C. J. van Rijsbergen, & P. W. Williams (Eds.), *Information Retrieval Research, Proceedings of the Joint ACM/BCS Symposium in Information Storage and Retrieval, Cambridge, June 1980* (pp. 23–34).
16. Carson, C., Belongie, S., Greenspan, H., & Malik, J. (1997). Region-based image querying. In *Proceedings of the IEEE Workshop CVPR '97 Workshop on Content-Based Access of Image and Video Libraries, Puerto Rico* (pp. 42–49).
17. Singitham, P. K. C., Mahabhashyam, M. S., & Raghavan, P. (2004). Efficiency-quality tradeoffs for vector score aggregation. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, K. B. Schiefer (Eds.), *Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, August 31–September 3 2004* (pp. 624–635). Burlington: Morgan Kaufmann.
18. Fuhr, N., & Großjohann, K. (2001). XIRQL: A query language for information retrieval in xml documents. In W. B. Croft, D. J. Harper, D. H. Kraft, J. Zobel (Eds.), *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9–13, 2001, New Orleans, LA* (pp. 172–180). New York: ACM.
19. Oracle Corporation. (1999). *Oracle8i visual information retrieval – user guide and reference*.

20. Dubois, D., & Prade, H. (1986). Weighted minimum and maximum operations in fuzzy set theory. *Information Science, 39*, 205–210
21. Pasi, G. (1999). A logical formulation of the Boolean model and of weighted Boolean models. In *Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems (LUMIS 99)*. London: University College London.
22. Yager, R. R. (1987). A note on weighted queries in information retrieval systems. *Journal of the American Society for Information Science, 38*, 23–24.
23. Boughanem, M., Loiseau, Y., & Prade, H. (2005). Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In M. Detyniecki, J. M. Jose, A. Nürnberger, C. J. van Rijsbergen (Eds.), *Adaptive Multimedia Retrieval: User, Context, and Feedback, Third International Workshop, AMR 2005, Glasgow, July 28–29, 2005, Revised Selected Papers. Lecture notes in computer science* (Vol. 3877, pp. 44–54). Berlin: Springer.
24. Zellhöfer, D. (2015). *A Preference-Based Relevance Feedback Approach for Polyrepresentative Multimedia Retrieval*. Doctoral thesis, BTU Cottbus-Senftenberg.

Searching for Information with Meet and Join Operators

Emanuele Di Buccio and Massimo Melucci

Abstract Information Retrieval (IR) is the complex of models, languages, and techniques aimed to retrieve documents containing information relevant to the user's information needs. Current retrieval technology requires a retrieval model for guaranteeing effective results. While all retrieval models for term-based search bring into play the Boolean logic of sets, a document collection can be searched by themes, instead of terms, using the logic of vector spaces, instead of sets. Indeed, vector spaces may generalize sets by breaking some laws of algebra of sets. The main aim of this chapter is to provide an overview of the state-of-the-art formalism used in IR and explain how the novel model based on themes defined as vector spaces and inspired by quantum operators, such as two lattice operators known as meet and join, can be built upon this formalism.

1 Introduction

When searching for information, the users of an IR system express their information needs through behavior (e.g., click-through activity) or queries (e.g., natural language phrases). By its nature, IR is inherently an interactive activity which is performed by a user accessing the collections managed by a system through very interactive devices immersed in context. Queries, which are the most used data for expressing information needs, are sentences expressed in a natural language, oftentimes very short (e.g., one word) or occasionally much longer (e.g., a text paragraph). However, queries are not the only means to communicate information needs. Other means such as click-through data can be observed during user–system interaction or within social networks. Through interaction, the user aims to refine his query, to provide additional evidence describing his information need or to indirectly tell his needs to the system.

E. Di Buccio · M. Melucci (✉)
Department of Information Engineering, University of Padova, Padova, Italy
e-mail: dibuccio@dei.unipd.it; massimo.melucci@unipd.it

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6_8

At the design level, current IR technology requires a retrieval model, that is, a set of algebraic structures to describe documents inspired by a mathematical theory, and a retrieval function mapping document and query representation to the numeric real field for measuring the degree to which a document contains information relevant to a query. The most effective models are currently based on Boolean logic, vector spaces, and probability theory of which the Binary Independence Retrieval (BIR) model, the Best Match N. 25 (BM25) extension and the language models are special cases. In addition to traditional models, some machine learning-based algorithms have been proposed to find the retrieval function by looking at the data collected during user–system interaction; for example, methods for learning to rank documents and neural network-based retrieval models have been quite successful for some years.

The mathematical model or the retrieval function, documents, and queries are mathematically represented as elements of *sets*, while the sets are labeled by *terms* or other document properties. It is a matter of fact that Boolean models by definition view terms as document sets and answer search queries with document sets obtained by set operators; moreover, the probabilistic models are all inspired by the Kolmogorov theory of probability, which is intimately related to set theory; finally, the models based on vector spaces will eventually be a means of providing a ranking or a measure for sets because they assign a weight to terms and then to documents in the sets labeled by the occurring terms. The implementation of content representation in terms of keywords and posting lists reflects the view of terms as sets of documents and the view of retrieval operations as set operators. In this chapter, we suggest that a document collection can be searched by *themes*, instead of terms, by using the ultimate logic of *vector subspaces*, instead of sets. The basic idea is that a theme corresponds to a vector space and the retrieval operations correspond to the vector space operators, such as the two lattice operators known as meet and join. The trace operator provides a mathematical description of a ranking function of vector spaces.

2 Background

In this section we provide a background of the three main theories, i.e., set theory, vector spaces, probability and the relationships thereof underlying the unifying framework advocated in [26]. In particular, we would like to emphasize how some elements of a theory reformulate some elements of another theory, thus pointing out resemblances, dissimilarities, and possibly latent qualities or abilities that may be developed and lead to future retrieval models.

2.1 Vector Spaces

This chapter utilizes the following definitions.

Definition 1 (Vector Space) A vector space V is a set of points called "vectors" subject to two conditions:

- the multiplication of a vector of the space by a constant of a field is a vector of the same space,
- the addition of two vectors of the space is a vector of the same space.

Definition 2 (Linear Independence) A set of d vectors $|v_1\rangle, \ldots, |v_d\rangle$ of V are independent when

$$c_1 |v_1\rangle + \cdots + c_d |v_d\rangle = 0 \tag{1}$$

only if

$$c_i = 0 \qquad \forall i = 1, \ldots, d,$$

that is, no vector of the set is a linear combination of the other vectors of the set.

Definition 3 (Basis) A basis is a set of linearly independent vectors.

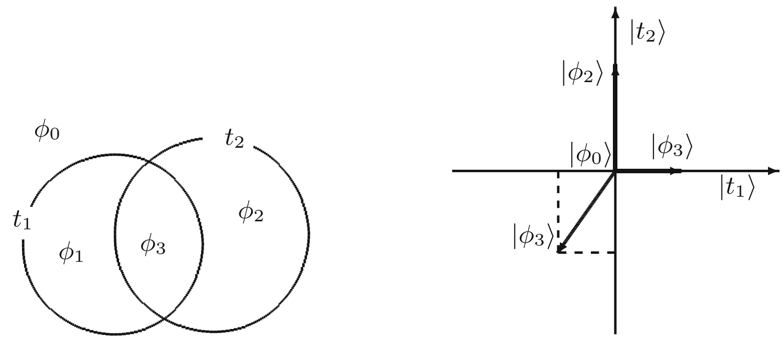
Definition 4 (Dimension) The dimension of V is d.

2.2 *Sets* Versus *Vector Spaces*

First of all, consider the foundational differences between a set and a space. A set is a primitive collection in which the elements cannot be combined together [11], whereas a space is a set in which the points, i.e., the vectors, can be mathematically combined to obtain other points of the same space [12]. Figure 1 depicts a collection of four documents indexed by two terms by using a twofold representation: one representation is based on vector spaces and the other representation is based on sets. The figure makes a basic difference between sets and vector spaces very clear; the latter allows the documents and the terms to relate through linear transformation combining and rotating one vector to another; the former does not allow any transformation and no element in a set is related with any other element.

It follows that a set is a more general concept of space, since a space is a set ruled out by mathematical mechanisms generating points; moreover, the points of a space are numbers or numerical tuples over a certain field, whereas the elements of a set can be of any kind.

Despite the differences, sets are related to spaces. The relationship between sets and spaces can be viewed through the notion of basis of an n-dimensional vector space V defined over a given field and provided with the inner product



(a) (b)

Fig. 1 On the right a bi-dimensional vector space is depicted as a Cartesian system. It includes four vectors or points. One vector coincides with the origin of the system; as both coordinates are null the vector corresponds to a document which is not indexed by any term. One vector lies on one axis and does not lie on the other axis; it corresponds to a document which is indexed by one term and is not indexed by the other. Another vector is similar to the previous one. Finally, one vector lies inside the plane and corresponds to a document which is indexed by both terms. On the left the four documents are placed in a Venn diagram where the sets are labeled by the terms and the elements correspond to the documents. (**a**) Representation based on sets. (**b**) Representation based on vector spaces

operator $\langle x|y\rangle$ for every pair of vectors $x, y \in V$. Consider the orthonormal basis $|t_1\rangle, \ldots, |t_n\rangle$ such that

$$\langle t_i|t_j\rangle = 1 \quad i = j \qquad \langle t_i|t_j\rangle = 0 \quad i \neq j. \tag{2}$$

Each $|t\rangle \in V$ corresponds to a property of an element. Each element is then assigned a vector of V as follows:

$$|v\rangle = x_1 |t_1\rangle + \cdots + x_n |t_n\rangle, \tag{3}$$

where $x_i \neq 0$ when the property t_i holds in the element v otherwise $x_i = 0$; in other words, term t_i is a feature of v when the basis vector $|t_i\rangle$ participates in the definition of $|v\rangle$. A special kind of basis is the canonical set of vectors such that

$$\langle t_i| = \begin{matrix} (\ 0, & \ldots, & 1, & \ldots, & 0\ ) \\ 1 & \ldots & i & \ldots & n \end{matrix}, \tag{4}$$

thus making $|v\rangle = |x\rangle$.

2.3 The Boolean Model for IR

The Boolean model for IR is by definition based on sets. According to the Boolean model, a content descriptor corresponds to a set of documents and then a document corresponds to an element of a set; for example, an index term corresponds to the set of documents indexed by that term. The Boolean model views a query as a Boolean expression in which the index terms are the operands and the operators are the usual disjunction, conjunction, and negation operators. In general, a query y can be written as the following conjunctive normal form:

$$y_1 \cap \cdots \cap y_n,$$

where the y_i's are conjoined propositions and the $y_{i,j}$'s are disjoined propositions, i.e.,

$$y_i = y_{i,1} \cup \cdots \cup y_{i,n_i}.$$

The documents managed by a system based on the Boolean model can be either retrieved or missed when they are matched against a query; therefore, the outcome of the system is binary and no ranking nor ordering is provided. To overcome this limitation, the coordination level assigns a measure to the operators and to any Boolean expression; therefore, the coordination level adds a score to every retrieved document and provides a ranking of a document list; for example, a certain weight function, such as max, is applied to each disjunction y_i to obtain the weight of y_i and then another weight function, such as the sum, can be applied to the scores given to y_i for all i, thus obtaining a score for y.

The Boolean model has been quite popular among expert users of early IR systems. If properly operated, a system can effectively retrieve both large proportions of documents relevant to many information needs and small proportions of non-relevant documents. The effectiveness of the Boolean model is due to the very natural view of a document collection as a set of documents. Thanks to this view, a user expects to receive a set of documents as the answer to his query.

Although the users of the World Wide Web (WWW) search engines are mostly reluctant to adopt Boolean operators as they are perceived to be bewildering, the search engines automatically insert disjunctions and conjunctions depending on the number of retrievable documents and by using some coordination level functions. Any other language, such as the Query-by-Theme Language (QTL) introduced in this chapter, might be perceived as much more complex than the Boolean language and it will very likely be perceived as cumbersome by many users; therefore, the QTL should be considered as a means for the retrieval system to operate on the user's input and then provide an alternative document ranking.

2.4 The Vector Space Model (VSM)

The VSM deviates from the naïve set theory and equips sets with linear relationships among vectors representing documents and queries. One net result is the provision of a principled mechanism to link documents, queries, and other retrieval constructs to the algebra of vector spaces. Flexibility and ease of application has been the main strengths and reasons of industrial and scientific success of the VSM. The vectors represent the occurrence of a term in a document or query; for example,

- if one term were available, each document would be associated with a number, that is, a point in a one-dimensional vector space which is geometrically depicted as a ray; the aforementioned scalar would correspond to the weight of the term in a document; if the document were an element of a set, no weight could be assigned because an element occurs once in a set unless such a Boolean view is provided with coordination level;
- if two terms were available, each document would be associated with two numbers, that is, it would be a point in a bi-dimensional vector space, which is geometrically a plane, where each vector component, e.g., $(0, 1) \in \mathbb{R}^2$, would denote the occurrence of one distinct term in a document in such a way that term 1 does not occur, while term 2 occurs in a document or in a query;
- if three terms were available, each document would be associated with three numbers, that is, it would be a point in a tridimensional vector space, which is geometrically a cube, where each vector component, e.g., (0, 1, 1), would denote the occurrence of one distinct term in a document in such a way that term 1 does not occur, while terms 2 and 3 occur in a document or in a query.

The vectorial representation of documents and queries implicitly assumes one orthonormal basis $|t_1\rangle, \ldots, |t_n\rangle$ such that each $|t\rangle$ corresponds to a term and each vector corresponds to a document or a query. The basis plays a crucial role since it defines a set of projectors where each projector is a binary function providing information about the occurrence of the term. Given a term t, the function $|t_i\rangle\langle t_i|$ is a projector such that $\langle t_i|t_i\rangle \langle t_i|t_i\rangle = \| \langle t_i|t_i\rangle \|^2 = 1$ and $\langle t_j|t_i\rangle \langle t_i|t_j\rangle = 0$ for any $|t_j\rangle \neq |t_i\rangle$.

The inner product between the query vector and the document vector becomes a principled explanation of the coordination level and becomes the retrieval function of the VSM. In terms of set theory, the inner product between a document vector and a query vector is a principled version of the coordination level since it can be viewed as the sum of the weights of the document memberships to the sets to which the query belongs. More specifically, the document vector can be expressed as

$$|x\rangle = x_1 |t_1\rangle + \cdots + x_n |t_n\rangle \tag{5}$$

and the query vector can be expressed as

$$|y\rangle = y_1 |t_1\rangle + \cdots + y_n |t_n\rangle . \tag{6}$$

The inner product can be written as

$$\langle x|y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i\, y_j\, \langle t_i|t_j\rangle\,, \tag{7}$$

where $\langle t_i|t_j\rangle = 1$ if and only if both document and query belong to the set corresponding to the intersection of sets t_i and t_j.

The choice of a function that assigns a weight (e.g., x_i or y_j) can only be empirically selected. To the aim of finding the best weight function, a series of experiments led to the conclusion that some weight functions such as Term Frequency (TF) $\times$ Inverse Document Frequency (IDF) (TFIDF) can be more effective than others for the most part [21]. The weighting schemes utilized by the retrieval functions of the VSM are perhaps the most important component of a retrieval system. Indeed, the occurrence of terms in documents is insufficient to achieve high levels of effectiveness. In mathematical terms, the strength of expressions like (5) and (6) is provided by the x's and the y's rather than by the basis vectors $|t\rangle$ and the ability of inner products like (7) to approximate relevance lies in the products $x\ y$ since the inner products $\langle t_i|t_j\rangle$ are trivially either 0 or 1.

A comparison between the notation of a model based on sets and the notation of a model based on vector spaces is summarized in Table 1 which introduces the notion of projector and space, since each space corresponds to one and only one projector. The analogous correspondence between sets and weight function does not exist. The strength of the correspondence between spaces and the projector is that the latter can be represented by a matrix and it thus provides an algorithmic implementation of checking whether a vector belongs in a space; it is indeed sufficient to compute two inner products and check whether the result is 1. Thanks to the correspondence between projector and subspace, a space of H can be viewed as a set of vectors where the projector plays the role of the mechanism that checks whether a vector belongs to the subspace.

The traditional VSM for IR ignores lattice operators like meet and join; it only exploits inner products and represents documents and queries by using only one basis unless Latent Semantic Analysis (LSA) is utilized. One reason for this limitation might be due to the greater focus of VSM-based system designers on (a) the least expert end users than on the users who are expert in their specific knowledge domain, on the one hand, and (b) the simple and short queries submitted

Table 1 Comparison between sets and vector spaces is summarized

Set	S	Vector space	H
Subset	a	Subspace	a
Set element	x	Vector	$\|x\rangle$
Weight function	W	Projector	$\mathbf{A}$
Ranking function	$W(x, a)$	Projection size	$\langle x\|\mathbf{A}\|x\rangle$
Membership	$W(x, a) = 1$	Projection	$\langle x\|\mathbf{A}\|x\rangle = 1$

to find specific resources, on the other hand. Although the least expert users would perceive little benefit from advanced vector operators, an IR may still be equipped with algorithms and data structures implementing these operators.

2.5 The Probabilistic Models

The role played by probabilistic models has become important in IR since the Boolean model lacks ranking capabilities and the end user has to face null output and output overload. The VSM succeeded in improving the user's experience because it provides some rankings, but finding the coefficients of the linear combinations has been an open problem for a long time and was mostly addressed through empirical investigations.

While weights are oftentimes provided by empirical investigations within the VSM, to the contrary, the probabilistic models provide weight functions with a sound theoretical basis such as Maximum Likelihood Estimation (MLE). A probabilistic model is currently a principled approach for providing the coordination level weights of which the BM25 is the most striking example. For instance, the so-called BIR owes its effectiveness to the Term Relevance Weight (TRW) function, which is a log-likelihood ratio from which BM25 was derived [20]. Statistical independence was further addressed by many authors, for example, in [6]. Similar and additional weight functions can be derived within the Language Modelling (LM) framework [7].

The probabilistic models organize an application domain as sets of single occurrences—elementary events—of a process or phenomenon. Elementary events are then grouped in subsets through random variables and a probability measure maps a subset of events, i.e., random values, to the unit range [0, 1] in order to obtain a probability. In general, the elementary events are documents and the events correspond to logical combinations of terms, which are sets.

Suppose we are given n terms. There are 2^n combinations of term occurrences, each corresponding to a subset of documents. Let x be one of these subsets. The probability $p(x) = P(d \in x)$ that a relevant document d belongs to x can be estimated under the assumption of conditional independence of term occurrence, thus providing that

$$p(x) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{1-x_i}, \tag{8}$$

where $x_i \in \{0, 1\}$ denotes the occurrence of term t_i and p is the probability that t_i occurs in a relevant document.

Suppose that not only is occurrence observed, but a random variable $S_i(d) \in [0, 1]$ is also measured for each term t_i and document d. In this context, x is a n-dimensional subset of $[0, 1]^n$. A probability distribution of $S_i(d)$ can thus be defined as follows:

$$B(s_i(d))^{-1}\ p_i^{s_i(d)}\,(1-p_i)^{1-s_i(d)} \qquad B(s) = \text{beta}\,(1-s, s+1) - \text{beta}\,(1-s, s+2)\,. \tag{9}$$

Consider the probability distribution of $S_i(d)$ when d is non-relevant:

$$B(s_i(d))^{-1}\ q_i^{s_i(d)}\,(1-q_i)^{1-s_i(d)}. \tag{10}$$

The log-likelihood of the hypothesis testing relevance *versus* non-relevance is

$$\log \frac{P(d \in x | d \text{ is relevant})}{P(d \in x | d \text{ is not relevant})} = \sum_{i=1}^{n} s_i(d)\ \log \frac{p_i\,(1-q_i)}{q_i\,(1-p_i)} \tag{11}$$

which is actually the BM25 scoring of d when $s_i(d)$ is the saturation of t_i in d. The advent of BM25 and the effective term weighting scheme thereof have made probabilistic models the state of the art.

Even though logic, vectorial and probabilistic approaches are three pillars of IR modeling, a strong relationship exists between them. In summary:

– The Boolean logic model views documents and queries as members of sets corresponding to terms. The Boolean operators allow the end user to compose complex queries and express more elaborate concepts than those expressed by terms.
– The VSM ensures that terms correspond to basis vectors and adds the inner product between the vectors representing the sets of the Boolean model to provide a ranking function of the documents with respect to a certain set of query terms.
– The BM25 scoring enriches the inner product with weights given by the MLE of the p and q parameters of a Beta-like probability function of the saturation factor.

3 Meet and Join

Not only can projectors be combined as explained in Sect. 2, but they can also be combined by operators called *meet* and *join* which significantly differ from the traditional set operators implemented by projectors. Consider a vector space V and two subspaces U, W thereof; we have the following definitions.

Definition 5 (Meet) The meet of U and W is the largest subspace included by both U and W.

Definition 6 (Join) The join of U and W is the smallest subspace including both U and W.

Meet and join only resembles intersection and union of sets. In fact, some properties of set operators cannot hold for meet and join anymore; for example, the distributive law holds for sets, but it does not for vector spaces.

From the point of view of information access, an interpretation of meet and join is necessary. The interpretation provided in this chapter for these two operators starts from the notion of basis. A basis vector mathematically represents a basic concept corresponding to data such as keywords or terms. In the event of a canonical basis, the basis vectors represent the starting vocabulary through which the content of documents and queries is produced and matched. When a document and a query or two documents share a concept their respective mathematical representations share a basis vector with non-null weight.

Consider the meet of two planes. The result of meeting two distinct planes is a ray, that is, a one-dimensional subspace. A one-dimensional subspace is spanned by a vector. Any vector can belong to any basis; indeed, the vector spanning the ray is the only vector of the basis of this subspace. As a basis vector can be a mathematical representation of a basic concept, the meet of two planes can be a mathematical representation of a basic concept. The planes meeting at the basis vector represent information sharing one concept, i.e., the concept represented by the meet, since the vector resulting from the meet of two planes may be a basis vector for both planes provided that each plane is spanned by a basis including the meet and another independent vector.

Consider the join of two distinct rays. The result of joining two rays is a plane, that is, a bi-dimensional subspace is spanned by two vectors. The subspace resulting from joining two rays is spanned by the vectors spanning the rays. The plane resulting from the join of two rays represents information based on two concepts, i.e., the concept represented by the basis vector of one ray and the concept represented by the basis vector of the other ray. Indeed, the vectors belonging to the plane resulting from the join of two rays are expressed by two basis vectors, each basis vector representing one individual concept.

However, it is safe to state that meet and join are only a mathematical representation and nothing can be argued about the meaning of what these two operators represent; we can nevertheless argue that if the planes meeting at the basis vector or the rays joined to a plane represent information sharing one concept or consisting of two concepts, respectively, the vector resulting from the meet of the two planes or the basis resulting from the join of two rays may be viewed as a sensible mathematical representation of complex concepts.

4 Structures of a Query-by-Theme Language

This section introduces the building blocks of a QTL. First, features and terms are introduced in Sect. 4.1. Then, Sect. 4.2 presents themes that are further exploited to rank documents as explained in Sect. 4.3. Finally, the composition of themes by

Table 2 Notations used in this chapter

Symbol	Meaning	Comment
$\|w\rangle$	Feature	Textual keyword and other modality depending on media
$\|t\rangle$	Term	Unigrams, bigrams, trigrams, etc., such as `information retrieval` and `quantum theory`
$\|\tau\rangle$	Theme	Expressions like `information retrieval` $\wedge$ `quantum theory` or `information retrieval` $\vee$ `quantum theory`
$\|\phi\rangle$	Document	Webpages, news, posts, etc.

using meet and join is described in Sect. 4.4. Table 2 summarizes the notation used in this chapter.

4.1 Features and Terms

Consider the features extracted from a collection of documents; for example, a word is a textual feature, the gray level of a pixel or a code word of an image is a visual feature, and a chroma-based descriptor for content-based music representation is an audio feature. Despite their differences, the features extracted from a collection of multimedia or multimodal documents can co-exist together in the same vector space if each feature is represented by a canonical basis vector. Consider k distinct features and the following.

Definition 7 (Term) Given the canonical basis[1] $|e_1\rangle, \ldots |e_k\rangle$ of a subspace over the real field, a *term* is defined as

$$|t\rangle = \sum_{i=1}^{k} a_i \; |e_i\rangle = (a_1, \ldots, a_k)' \qquad a_i \in \mathbb{R},$$

where the a's are the coefficients with respect to the basis. Therefore, terms are a combination of features; for example, if $k = 2$ textual features, say "information" and "retrieval," then "information retrieval" is a term represented by

$$|\text{information retrieval}\rangle = a_{\text{information}} \; |\text{information}\rangle + a_{\text{retrieval}} \; |\text{retrieval}\rangle \, .$$

The main difference between features and terms lies in orthogonality, since the feature vectors assume mutual orthogonality whereas the term vectors only assume mutual independence. Non-orthogonal independence also distinguishes the QTL from the VSM, since term vectors might not be—and they are often not—orthogonal whereas keyword vectors are usually assumed orthogonal; for example,

[1]The i-th canonical basis vector has $k - 1$ zeros and 1 at the i-th component.

$$|\text{retrieval system}\rangle = a_{\text{system}} \; |\text{system}\rangle + a_{\text{retrieval}} \; |\text{retrieval}\rangle$$

is not orthogonal to, yet it is still independent of $|\text{information retrieval}\rangle$.

4.2 Themes

Consider a vector space over the real field and the following:[2]

Definition 8 (Theme) Given m independent term vectors $|t_1\rangle, \dots, |t_m\rangle$, where $1 \le m \le k$, a *theme* is represented by the m-dimensional subspace of all vectors like

$$|\tau\rangle = b_1 \; |t_1\rangle + \cdots + b_m \; |t_m\rangle \qquad b_i \in \mathbb{R}.$$

From the definition, one can see that a feature is a term and a term is the simplest form of a theme. In particular, a term is a one-dimensional theme. Moreover, if $|t\rangle$ is a term, then any $c \; |t\rangle$ is the same term for all $c \in \mathbb{R}$.

Moreover, themes can be combined to further define more complex themes; to start with, a theme can be represented by a one-dimensional subspace (i.e., a ray) as follows: if $|t\rangle$ represents a term, we have that $|\tau\rangle = b \; |t\rangle$ spans a one-dimensional subspace (i.e., a ray) and represents a theme. Also, a theme can be represented by a bi-dimensional subspace (i.e., a plane) in the k-dimensional space as follows: if $|t_1\rangle$ and $|t_2\rangle$ are term vectors, we have that $b_1 \; |t_1\rangle + b_2 \; |t_2\rangle$ spans a bi-dimensional subspace (i.e., a plane) representing a theme.

In general, a theme may be represented by multi-dimensional subspaces by using different methods; for example, "information retrieval systems" can be a term represented as a linear combination of three feature vectors (e.g., keywords) or it can be a theme represented by a linear combination of two or more term vectors such as "information retrieval," "retrieval," "systems," "retrieval systems," or "information." Therefore, the correspondence between themes and subspaces is more complex than the correspondence between keywords and vectors of the VSM. The conceptual relationships between themes are depicted in Fig. 2.

4.3 Document Ranking

A document is represented by a vector

$$|\phi\rangle = (c_1, \dots, c_m)' \qquad c_i \in \mathbb{R}$$

[2]Note that we are overloading the symbol $|\tau\rangle$ to mean both the theme subspace and a vector of that subspace.

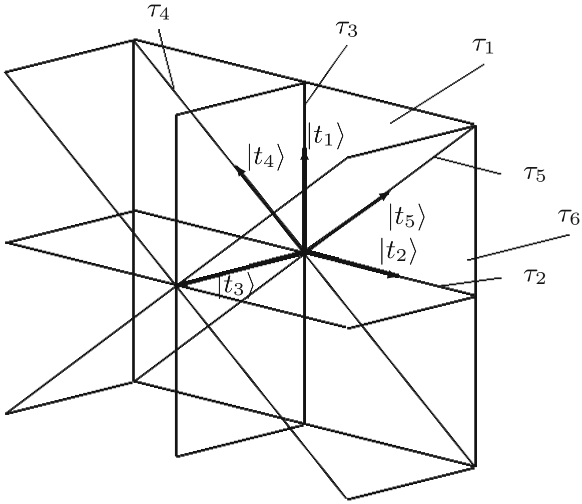


Fig. 2 A pictorial representation of features, terms, and themes. Three feature vectors $|e_1\rangle$, $|e_2\rangle$, $|e_3\rangle$ span a tridimensional vector space, but they are not depicted for the sake of clarity; the reader may suppose that the feature vectors are any triple of independent vectors. Each term vector $|t\rangle$ can be spanned by any subset of feature vectors; for example, $|t_1\rangle = a_{1,1}\ |e_1\rangle + a_{1,2}\ |e_2\rangle$ for some $a_{1,1}, a_{1,2}$. A theme can be represented by a subspace spanned by term vectors; for example, $|t_1\rangle$ and $|t_2\rangle$ span a bi-dimensional subspace representing a theme and including all vectors $|\tau_1\rangle = b_{1,1}\ |t_1\rangle + b_{1,2}\ |t_2\rangle$

on the same basis as that which is used for themes and terms such that c_i is the measure of the degree to which the document represented by the vector is about term i.

The ranking rule for measuring the degree to which a document is about a theme relies on the theory of abstract vector spaces. To measure this degree, a representation of a document in a vector space and a representation of a theme in the same space are necessary. Document and theme share the same representation, if they are expressed with respect to the same basis of m term vectors. When orthogonality holds, a ranking rule is then the squared projection of $|\phi\rangle$ on the subspace spanned by a set of m term vectors as explained in [17].

To describe the implementation of the ranking rule, projectors are necessary. To this end, an orthogonal basis of the same subspace can be obtained through linear transformation of the $|t\rangle$'s. Let $\{|v_1\rangle, \ldots, |v_m\rangle\}$ be such an orthogonal basis, which determines the projector of the subspace as follows:

$$\boldsymbol{\tau} = |v_1\rangle\langle v_1| + \cdots + |v_m\rangle\langle v_m|.$$

The measure of the degree to which a document is about a theme τ represented by the subspace spanned by the basis vectors $|v\rangle$ is the size of the projection of the document vector on the theme subspace, that is,

$$\mathrm{tr}[\boldsymbol{\tau} \, |\phi\rangle\langle\phi|] = \langle\phi| \, \boldsymbol{\tau} \, |\phi\rangle \,, \tag{12}$$

where tr is the trace operator. After a few passages, the following measure is obtained by leveraging orthogonality [11]:

$$|\,\langle v_1|\phi\rangle\,|^2 + \cdots + |\,\langle v_m|\phi\rangle\,|^2. \tag{13}$$

4.4 Meet and Join Operators

Themes can be created through operators applied to other themes defined on a vector space. In this chapter, we introduce two operators called *meet* and *join*. Thus, the subspaces that represent a theme can meet or join the subspace of another theme and the result of either operation is a subspace that represents yet another theme.

The intuition behind using meet and join in IR is that in order to significantly improve retrieval effectiveness, users need a radically different approach to searching a document collection that goes beyond the classical mechanics of an IR system; for example, the distributive law of intersection and union does not remain valid for subspaces equipped with meet and join. Although it is a negative feature of a classical theory, the violation of a property can be a potential advantage of QTL since this violation allows a user who is interacting with a retrieval system to experiment with many more expressions of his information need. First, consider the following definition of join.

Definition 9 (Join) Consider r themes $\tau_1, \ldots, \tau_r$. Each theme τ_i corresponds to one subspace spanned by a basis $|t_{i,1}\rangle, \ldots, |t_{i,k_i}\rangle$, where k_i is the dimension of the i-th subspace. The join of r themes can be defined by

$$\tau_1 \vee \cdots \vee \tau_r$$

and includes the vectors resulting from

$$b_{1,1}\;|t_{1,1}\rangle + \cdots + b_{1,k_1}\;|t_{1,k_1}\rangle + \cdots + b_{n,1}\;|t_{n,1}\rangle + \cdots + b_{r,k_r}\;|t_{r,k_r}\rangle\,.$$

In the event of $r = 2, k_1 = 1, k_2 = 1$, two rays are joined, thus resulting in a plane; see Fig. 3a. Note that the join is the smallest subspace containing all the joined subspaces. Then, consider also the following definition of meet.

Definition 10 (Meet) Consider t themes $\tau_1, \ldots, \tau_t$ of dimension $k_1, \ldots, k_t$. Each theme i corresponds to one subspace spanned by a basis $|t_{i,1}\rangle, \ldots, |t_{i,t_i}\rangle$. The meet of t themes can be defined by

$$\tau_1 \wedge \cdots \wedge \tau_t$$

and includes the vectors resulting from

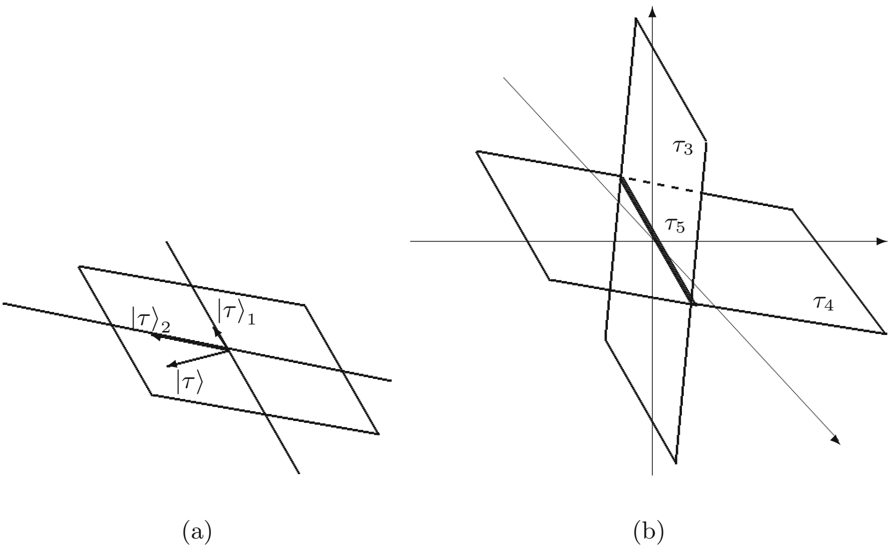


Fig. 3 Pictorial description of join and meet. (**a**) Join. (**b**) Meet

$$a_1 \; |v_1\rangle + \cdots + a_{\min k_1, k_t} \; |v_{\min k_1, k_t}\rangle \,,$$

where the $|v\rangle$'s are the basis vectors of the largest subspace contained by all subspaces.

In the event that $t = 2$, $k_1 = 2$, and $k_2 = 2$, the meet may result in one ray, i.e., the intersection between two planes; see Fig. 3b.

In general, the distributive law is violated by themes. For all $\tau_1, \tau_2, \tau_3, \tau_4$ we have that

$$(\tau_1 \wedge \tau_2) \vee (\tau_1 \wedge \tau_3) \neq \tau_1 \wedge (\tau_2 \vee \tau_3) \,;$$

therefore,

$$(\tau_1 \wedge \tau_2) \vee (\tau_1 \wedge \tau_3)$$

calculates one ranking, while

$$\tau_1 \wedge (\tau_2 \vee \tau_3)$$

might yield another ranking, thus giving the chance that one ranking is more effective than another ranking, that is, the lack of the distributive property gives one further degree of freedom in building new information need representations. The violation of the distributive property is shown as follows: Fig. 2 shows three term vectors, i.e., $|t_1\rangle$, $|t_2\rangle$, and $|t_3\rangle$, spanning a tridimensional vector space; each of these term vectors spans a one-dimensional subspace, i.e., a ray. Note that the bi-dimensional subspace, i.e., a plane, spanned by $|t_1\rangle$ and $|t_2\rangle$ is also spanned by $|t_4\rangle$ and $|t_5\rangle$. Following the explanation of [14] and [26, pp. 38–39], consider the subspace spanned by

$$t_2 \wedge (t_4 \vee t_5).$$

As the bi-dimensional subspace spanned by $|t_1\rangle$ and $|t_2\rangle$ is also spanned by $|t_4\rangle$ and $|t_5\rangle$ we have that

$$t_2 \wedge (t_4 \vee t_5) = t_2 \wedge (t_1 \vee t_2) = t_2.$$

Let's distribute meet. We have that

$$(t_2 \wedge t_5) \vee (t_2 \wedge t_4) = \emptyset$$

because

$$t_2 \wedge t_5 = \emptyset \qquad t_2 \wedge t_4 = \emptyset.$$

Therefore ,

$$t_2 = t_2 \wedge (t_4 \vee t_5) \neq (t_2 \wedge t_4) \vee (t_2 \wedge t_5) = \emptyset$$

thus meaning that the distributive law does not hold; hence, set operations cannot be applied to subspaces.

5 Implementation of a Query-by-Theme Language

Given m terms, k features, and n documents, a $k \times n$ matrix $\mathbf{X}$ can be computed such that $\mathbf{X}[i, j]$ is the frequency of feature i in document j; frequency is only one option, but $\mathbf{X}$ may contain other non-negative weights. As $\mathbf{X}$ is non-negative, Non-negative Matrix Factorization (NMF) [16] can be performed in such a way to obtain:

$$\mathbf{X} = \mathbf{W}\,\mathbf{H} \qquad \mathbf{W} \in \mathbb{R}^{k \times m} \qquad \mathbf{H} \in \mathbb{R}^{m \times n}, \tag{14}$$

where $\mathbf{H}[h, j]$ measures the contribution of theme h to document j. As the themes are unknown, they have to be calculated as follows. The m column vectors of $\mathbf{W}$ are regarded as terms, i.e., one-dimensional themes. The theme vectors corresponding to the columns of $\mathbf{W}$ are then rescaled as follows:

$$(|\tau_1\rangle, \ldots, |\tau_m\rangle) = \mathbf{W}\,\mathrm{diag}(\mathbf{H}\,1_n),$$

where 1_n is the vector of n 1's and "diag" transforms a vector into a diagonal matrix. In this way, each element i of every column vector of $\mathbf{W}$ is multiplied by the sum of the coefficients of row i of $\mathbf{H}$; as $\mathbf{H}[h, j]$ measures the contribution of theme h to document j, this multiplication multiplies element i of each column vector of $\mathbf{W}$ by the total contribution of term i to themes.

The definition of join and meet requires algorithms for computing an effective representation of the subspaces stemming from these operators. To this end, we consider the following:

1. the join of two one-dimensional themes, and
2. the meet of two bi-dimensional themes.

We limited ourselves to the bi-dimensional case for the sake of simplicity. The join algorithm consists of rotating two theme vectors $|\tau_1\rangle$, $|\tau_2\rangle$ to obtain $|u_1\rangle$, $|u_2\rangle$ as depicted in Fig. 4a. The implementation consists of the JOIN function as follows:

1. The function is called with two one-dimensional subspaces as parameters.
2–4. The passed parameters are normalized so that the ℓ_p-norm is one ($p = 2$).
5. One real coefficient is the inner product value between the passed parameters; it will be used at step 8. This coefficient may also be viewed as the quantum probability that one parameter is the same as the other because its square lies between zero and one.
6. The other real coefficient is the complement of the first coefficient; it will be used at step 8. It can also be viewed as the complement quantum probability.
7. The first output vector is the first parameter.
8. The second output vector results from a rotation.
9. The output is an orthogonal basis of the plane spanned by the parameters.

The meet algorithm for two bi-dimensional themes consists of (1) the algorithm of the join for obtaining the representation of each bi-dimensional subspace and (2) the algorithm for calculating the solution of the linear system

1: JOIN(τ_1, τ_2)
2: **for all** $i = 1, 2$ **do**
3: $\quad |\tau_i\rangle \leftarrow |\tau_i\rangle \,/\, \sqrt{\langle\tau_i|\tau_i\rangle}$
4: **end for**
5: $a_2 \leftarrow \langle\tau_1|\tau_2\rangle$
6: $b_2 \leftarrow \sqrt{1 - a_2^2}$
7: $|u_1\rangle \leftarrow |\tau_1\rangle$
8: $|u_2\rangle \leftarrow (|\tau_2\rangle - a_2\,|\tau_1\rangle)/b_2$
9: **return** $|u_1\rangle, |u_2\rangle$

(a)

1: MEET$(\tau_1, \tau_2, \tau_3, \tau_5)$
2: $|u_1\rangle, |u_2\rangle \leftarrow$ JOIN(τ_1, τ_2)
3: $|u_3\rangle, |u_4\rangle \leftarrow$ JOIN(τ_3, τ_4)
4: $\mathbf{A} \leftarrow (|u_1\rangle, |u_2\rangle, |-u_3\rangle)$
5: $\mathbf{Q}, \mathbf{R} \leftarrow \mathrm{QR}(\mathbf{A})$
6: $|q_b\rangle \leftarrow \mathbf{Q}\,|u_4\rangle$
7: $|x\rangle \leftarrow$ solution of $\mathbf{R}\,|x\rangle = |q_b\rangle$
8: $|v\rangle \leftarrow x_1\,|u_1\rangle + x_2\,|u_2\rangle$
9: **return** $|v\rangle$

(b)

Fig. 4 Efficient computation of meet and join; the join algorithm is inspired by Gram–Schmidt's procedure. (**a**) The join algorithm. (**b**) The meet algorithm

$$c_1 \ |u_1\rangle + c_2 \ |u_2\rangle = c_3 \ |u_3\rangle + c_4 \ |u_4\rangle \ ,$$

where $\{ |u_1\rangle \ , \ |u_2\rangle\}$ is the basis of one bi-dimensional subspace and $\{ |u_3\rangle \ , \ |u_4\rangle\}$ is the basis of the other bi-dimensional subspace as described in Fig. 4b. The implementation consists of the MEET function as follows:

1. The function is called by any other function and four one-dimensional subspaces are passed as parameters. The first two τ's are the basis vectors of one plane.
2. The orthogonal basis of the first plane is calculated by the join of $|\tau_1\rangle$ and $|\tau_2\rangle$.
3. The orthogonal basis of the second plane is calculated by the join of $|\tau_3\rangle$ and $|\tau_4\rangle$.
4. The special matrix **A** is built by simply aggregating three out of the four vectors calculated in the previous two steps.
5. The QR decomposition represents the same transformation as **A** whose columns are replaced by the orthonormal columns of **Q**. As **R** is triangular, computing a transformation is more efficient.
6. Indeed, this transformation maps the fourth plane vector into the constant vector of a linear system whose coefficient is **R**.

7–9. Finally, the two values of the solution of the aforementioned linear system are the coordinates of the meet with respect to the orthogonal basis vectors of the first plane.

6 Related Work

This chapter proposes a new paradigm to express the user information need through themes and provides a set of operators that the user can exploit to interact with those themes. To this aim, the extraction of themes as content complex descriptors of documents is a necessary step. The research on algorithms for extracting complex descriptors, e.g., a set of (possibly related) terms, on models for exploiting such descriptors for document ranking, and on approaches for interacting with those descriptors are all relevant to this chapter. Other research papers are somehow relevant to this chapter; however, an exhaustive survey of literature about every topic is infeasible and perhaps unnecessary, thanks to textbooks such as [8]. We will provide some pointers on the essential work.

Since the early stages of research in IR, Query Expansion (QE) has been the standard approach to supporting the end user during the interaction with the retrieval system in place of manual query modification. A number of techniques that obtain an alternative description of the user's information need have been experimented and surveyed in [5]. Relevance Feedback (RF) and in particular

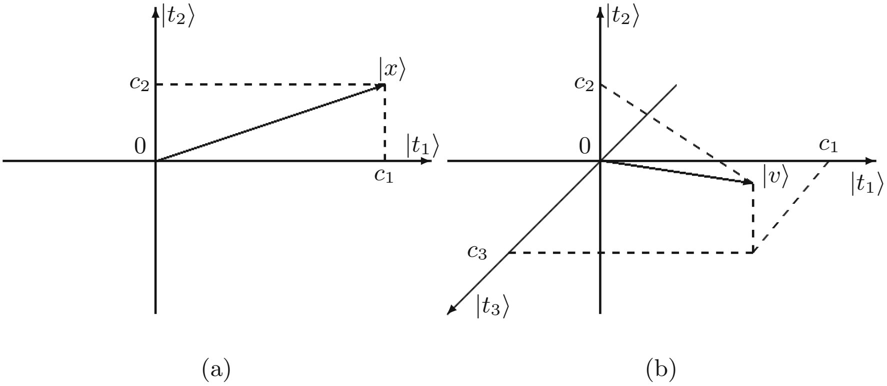


Fig. 5 How a user would build queries or documents by using the VSM

Pseudo Relevance Feedback (PRF)[3] have been a crucial propellant of QE and have essentially been based on matching queries and documents differently, though implemented according to a model.

As QTL exploited vector spaces, the comparison with the VSM is quite natural, but some fundamental differences exist. The intuition behind the VSM is illustrated in Fig. 5a. A user starts writing a query without any keyword in mind; the user's starting point corresponds to the origin (0, 0) of a coordinate system. Once the user has selected a keyword t_1, the point moves to $|t_1\rangle$ with weight or coordinate c_1. If the user is an author $c_1\ |t_1\rangle$ represents a document; if the user is a searcher $c_1\ |t_1\rangle$ represents a query. When the user chooses t_2 with weight or coordinate c_2, the query or document vector is $c_1\ |t_1\rangle + c_2\ |t_2\rangle$. When k keywords are selected, the final result is given by

$$c_1\ |t_1\rangle + \cdots + c_k\ |t_k\rangle\ .$$

The c's measure the importance of a keyword in the query or in the document. The same applies to multimedia or multimodal objects where the content descriptors can be video genre, music timbre, or color as depicted in Fig. 5b. Therefore, the rationale of the VSM differs from the rationale of the QTL, since our language clearly leverages the potential of the algebra of vector spaces, whereas the VSM limits itself to represent document and queries as vectors matched by means of inner products.

Another line of research that is relevant to the work reported in this chapter is automatic approaches to capture word relationships. LSA was proposed to extract

[3]"Pseudo" originates from Greek and means "falsehood"; when applied to feedback, "pseudo" means that relevance is not the true, real relevance provided by a user, on the contrary, is provided by a surrogate for the user, i.e., the system.

descriptors that capture word and document relationships within one single model [9]. In practice, LSA is an application of Singular Value Decomposition (SVD) to a document-term matrix. Following LSA, Latent Dirichlet Allocation (LDA) aims at automatically discovering the main topics in a document corpus. A corpus is usually modeled as a probability distribution over a shared set of topics; these topics in turn are probability distributions over words, and each word in a document is generated by the topics [2]. This chapter focuses on the geometry provided by vector spaces, yet is also linked to topic models, since a probability distribution over documents or features is defined in a vector space, the latter being a core concept of the quantum mechanical framework applied to IR [17, 18, 26].

The approaches to term dependency investigated in [23, 24] can supplement our QTL, even though those papers are focused on QE. Operators for vector spaces are mentioned in [4, 19], but meet and join are not explicitly modeled or implemented. In particular, both single terms and term dependencies can be modeled as elementary events in a vector space and dependencies can be modeled as superposition [24], interference [23], or tensor products [4, 19]. Moreover, in [19], rays describe information needs, which can be terms or features as well. These terms or features are combined using superposition or mixtures, for instance, but the authors do not explicitly use or evaluate quantum logics. Our contribution is the possibility that the user may explicitly add meet and join to the query, thus directly assessing the impact of the operators on retrieval results.

This chapter also provides an effective language to implement the principle of poly-representation [15], which aims to generate and exploit the cognitive overlap between different representations of documents to estimate an accurate representation of the usefulness of the document. Documents that lie in the same representations are relevant to a user's information need. Poly-representation was described within the quantum mechanical framework in [10]. Indeed, the quantum mechanical framework may describe various aspects of document representation within the same space: fusion of document content representations; temporal aspects and dynamic changes; document structure and layout.

Efforts that aimed to implement query languages equipped with operators over vector spaces were made and they resulted in Quantum Query Language (QQL) [22]. For example, `SELECT * FROM t WHERE x='b' OR x='c'` can be modeled by finding the sum $\mathbf{P}_{bc} = \mathbf{P}_b + \mathbf{P}_c$ of the mutually orthogonal projectors corresponding to the subspaces spanned by b and c and then computing $\langle\phi|\,\mathbf{P}_{bc}\,|\phi\rangle$. In [29] the combination of the dual approaches reported in [10] and [22] is mentioned but not addressed.

Widdows introduced a principled approach to negative RF in [27, 28]. According to him, a retrieval system adds term vectors to the query vector when a user adds terms to describe his information need. Suppose two query terms t_1, t_2 both describe the need and the user wants all and only the documents indexed by both terms. To this end, he will submit a query like t_1 AND t_2. Suppose the user no longer wants the documents about t_2. To this end, if t_2 are no longer describing the need, then t_1 AND NOT t_2 would be the right query. According to the VSM, the term vectors should be subtracted from the query vector. This subtraction is actually negative

RF; however, the negative RF of the VSM requires that the β parameters be defined precisely. Although the VSM specifies what to do with the vectors to implement negative feedback, it does not provide insights on how to define the parameters. In contrast, vector rotation specifies how to define these parameters.

Finally, the literature on efficient posting list processing is worth mentioning because the meet and join algorithms require some efficient implementation. The algorithms described in [3] and [25], for example, may be useful resources because they aim to retrieve the most likely relevant documents from the sets of documents associated with the query terms as fast as possible. This chapter concentrates on document modeling and user interaction level, since meet and join operates on abstract representations of document and terms. Nevertheless, the theme model and meet and join can still be implemented within an IR system, thus benefitting from the efficient solutions reported in recent literature [13].

7 Discussion and Future Work

Two difficulties tie IR to Quantum Mechanics (QM). On the one hand, in the IR field the peculiar difficulty faced by a retrieval system of precisely and exhaustively describing relevance only using data is well known; for example, neither a system nor a user can describe a relevant document using text even though it adds many keywords. A user cannot even precisely and exhaustively describe his own information need. The only thing a system can do is infer relevance by the document content, the user's request, and all the other sources of evidence. On the other hand, in QM, theory can only approximately describe and predict the microscopic and invisible world due to the fragile state of the particles and the inherent uncertainty of measurement; there is an unbridgeable gap, and it lies between the unknowable world of subatomic particles and the outcomes produced by the devices used for describing this world. The similarities among the gap between content and relevance thereof, the gap between request and information need thereof, and the gap between subatomic particles and observed quantities thereof were the reason why some researchers investigated the quantum mechanical framework in IR. The fundamental idea underlying this utilization was the potential offered by the quantum mechanical framework to predict the values which can only be observed in conditions of uncertainty [17].

The utilization of quantum structures in Computer Science is attracting much attention [1]. One of the most asked questions about the utilization of quantum structures in IR in particular and in Computer Science in general still remains and it is about its practical and theoretical impact. The QTL described in Sect. 4 suffers the same fate and some questions arise about the necessity and the utility of deploying quantum structures in IR. In this regard there are two main aspects: one aspect is mainly experimental and practical (i.e., are quantum structures improving effectiveness?), the other is mainly theoretical (i.e., what kind of concepts can be modeled by quantum structures?). While the practical impact is also a matter of experimentation, the theoretical impact has been addressed since the proposal of the Geometry of IR [26]. In this chapter we address both aspects.

With regard to the practical impact of the QTL, some types of user, who are experts in their own application domains such as journalists and scholars, may be willing to use meet and join for building complex queries and searching a document collection by themes rather than simple and short queries and finding specific resources. A user may meet and join subspaces in the context of vector spaces, instead of intersecting and complementing subsets. Although meet and join are well-known operators of quantum theory, we do not argue that documents and queries are quantum objects like subatomic particles. Instead, we are investigating whether the retrieval process involving expert users may exhibit some quantum-*like* behavior.

From the theoretical perspective there is a more profound reason suggesting the replacement of sets with spaces. Actually, an initial replacement took place with the advent of the VSM which views documents as points of a vector space and not only mere elements of Boolean sets. The main motivation driving from the Boolean sets to the vector spaces was the need of a retrieval function providing a ranking. The inner product of the VSM between document vectors and query vectors provides such a ranking because it sums up the weights of the memberships of a document and the sets to which a query belongs.

One future work will focus on media other than text, terms, and words and on modalities other than querying. Indeed, a term is bound to the easy recognition of terms in documents and to the user's intuition that a term corresponds to the set of documents about the concept represented by the term. When terms are combined by Boolean operators, a term has a semantics and the results, which are document sets, obtained by the operators are an extensional representation of a concept. A set-based approach to retrieval with image, video, sound, or multimedia documents is less natural than with textual documents. The content descriptors of image, video, or sound such as pixels, shapes, or chroma cannot be described by terms and the assumption that sets and set operators can express informative content does not seem as intuitive as for text. Similarly, multimodality fits less naturally with a set-based retrieval model. When click-through and user interaction data are collected, sets are not the most obvious representation of informative content. The reason is that the language of non-textual or multimodal traits is likely to describe individuals with a logic other than a classical logic.

To the end of experimenting different modalities, some experiments are underway by using the subtopics of the TREC 2010 Web Track Test Collection as themes.[4] Instead of implementing themes using index terms, we will implement themes using subtopics, which may be viewed as aspects of the main topic. The experiments will simulate a more interactive scenario than the scenario simulated in this chapter. A user will submit a query (i.e., the main topic) and the retrieval system will extract a set of pertinent themes. We will measure the effectiveness of

[4] http://trec.nist.gov/data/web/10/wt2010-topics.xml.

the ranked list obtained by the representation which will be based only on the query terms, of the list obtained by using all the distinct terms associated with the extracted themes or by using the themes built through join and meet. Further experiments will be carried out on the Dynamic Domain Track Test Collections. The goal of the Dynamic Domain Track is to "support research in dynamic, exploratory search of complex information domains."[5] The task is highly interactive and the interaction with the user is simulated through the Jig, which returns explicit judgments on the top five retrieved documents along with relevant passages in those documents. We will investigate the use of relevant passages as a source for implementing themes.

Acknowledgements This chapter is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321.

References

1. Aerts, D., Melucci, M., de Bianchi, M. S., Sozzo, S., & Veloz, T. (2018). Special issue: Quantum structures in computer science: Language, semantics, retrieval. *Theoretical Computer Science, 752*, 1–4.
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
3. Broder, A. Z., Carmel, D., Herscovici, M., Soffer, A., & Zien, J. (2003). Efficient query evaluation using a two-level retrieval process. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (pp. 426–434). New York, NY: ACM. http://doi.acm.org/10.1145/956863.956944.
4. Caputo, A., Piwowarski, B., & Lalmas, M. (2011). A query algebra for quantum information retrieval. In *Proceedings of the IIR Workshop*. http://ceur-ws.org/Vol-704/19.pdf.
5. Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys, 44*(1), 1–50. http://doi.acm.org/10.1145/2071389.2071390.
6. Cooper, W. (1995). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems, 13*(1), 100–111.
7. Croft, W., & Lafferty, J. (Eds.). (2003). *Language modeling for information retrieval*. Berlin: Springer.
8. Croft, W., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice*. Boston: Addison Wesley. http://ciir.cs.umass.edu/downloads/SEIRiP.pdf.
9. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science and Technology, 41*(6), 391–407.
10. Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., & van Rijsbergen, K. (2010). Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proceedings of IIiX* (pp. 115–124).
11. Halmos, P. (1987). *Finite-dimensional vector spaces. Undergraduate texts in mathematics*. New York, NY: Springer.
12. Halmos, P. R. (1960). *Naïve Set Theory*. New York, NY: D. Van Nostrand Company, Inc.

[5] http://trec-dd.org.

13. Hawking, D., Moffat, A., & Trotman, A. (2017). Efficiency in information retrieval: Introduction to special issue. *Information Retrieval Journal, 20*(3), 169–171. http://dx.doi.org/10.1007/s10791-017-9309-7.
14. Hughes, R. (1989). *The structure and interpretation of quantum mechanics*. Cambridge: Harvard University Press.
15. Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham Publishing.
16. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788–791.
17. Melucci, M. (2015). *Introduction to information retrieval and quantum mechanics*. Berlin: Springer.
18. Melucci, M., & van Rijsbergen, C. J. (2011). *Quantum mechanics and information retrieval* (Chap. 6, pp. 125–155). Berlin: Springer.
19. Piwowarski, B., Frommholz, I., Lalmas, M., & van Rijsbergen, C. J. (2010). What can quantum theory bring to information retrieval. In *Proceedings of CIKM* (pp. 59–68). New York, NY: ACM.
20. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389.
21. Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.
22. Schmitt, I. (2008). QQL: A DB&IR query language. *The VLDB Journal, 17*(1), 39–56. http://dx.doi.org/10.1007/s00778-007-0070-1.
23. Sordoni, A., He, J., & Nie, J. Y. (2013). Modeling latent topic interactions using quantum interference for information retrieval. In *Proceedings of CIKM* (pp. 1197–1200). http://dl.acm.org/citation.cfm?doid=2505515.2507854
24. Sordoni, A., Nie, J. Y., & Bengio, Y. (2013). Modeling term dependencies with quantum language models for IR. In *Proceedings of SIGIR* (pp. 653–662).
25. Turtle, H., & Flood, J. (1995). Query evaluation: Strategies and optimizations. *Information Processing & Management, 31*(6), 831–850. http://dx.doi.org/10.1016/0306-4573(95)00020-H.
26. Van Rijsbergen, C. J. (2004). *The Geometry of information retrieval*. Cambridge: Cambridge University Press.
27. Widdows, D. (2004). *Geometry and meaning*. Stanford, CA: CSLI Publications.
28. Widdows, D., & Peters, S. (2003). Word vectors and quantum logic: Experiments with negation and disjunction. In R. T. Oehrle & J. Rogers (Eds.), *Proceedings of the Mathematics of Language Conference* (Vols. 141–154).
29. Zellhöfer, D., Frommholz, I., Schmitt, I., Lalmas, M., & van Rijsbergen, K. (2011). Towards quantum-based DB+IR processing based on the principle of polyrepresentation. In *Proceedings of ECIR* (pp. 729–732). Berlin: Springer. http://dl.acm.org/citation.cfm?id=1996889.1996989.

Index

D. Aerts et al. (eds.), *Quantum-Like Models for Information Retrieval and Decision-Making*, STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, https://doi.org/10.1007/978-3-030-25913-6

D

E

F

G

H

I