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 $|v_1\rangle, \dots, |v_d\rangle$ of V are independent when

$$c_1 |v_1\rangle + \cdots + c_d |v_d\rangle = 0 \tag{1}$$

only if

$$c_i = 0 \qquad \forall i = 1, \dots, d,$$

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

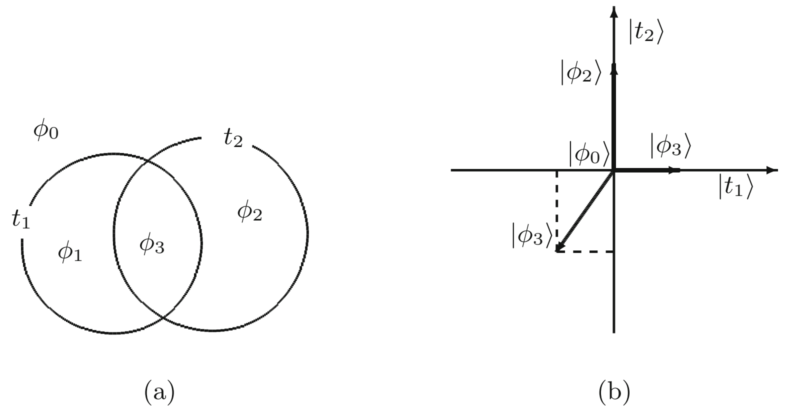


Fig. 1 On the right a bi-dimensional vector space is depicted as a Cartesian system. It includes four vectors or points. One vector coincides with the origin of the system; as both coordinates are null the vector corresponds to a document which is not indexed by any term. One vector lies on one axis 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 $|t_1\rangle, \dots, |t_n\rangle$ 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

$$\langle t_i| = \begin{array}{ccccccc} (\; 0, & \dots, & 1, & \dots, & 0 \;) \\ 1 & \dots & i & \dots & n \end{array}, \tag{4}$$

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) × 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

$$p(x) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{1-x_i}, \qquad (8)$$

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:

$$B(s_i(d))^{-1}\, p_i^{s_i(d)}\, (1-p_i)^{1-s_i(d)} \quad B(s) = \text{beta}\,(1-s, s+1) - \text{beta}\,(1-s, s+2)\,. \tag{9}$$

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 $|t_1\rangle, \ldots, |t_m\rangle$, 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

$$|\phi\rangle = (c_1, \ldots, c_m)' \qquad c_i \in \mathbb{R}$$

[2] Note that we are overloading the symbol $|\tau\rangle$ to mean both the theme subspace and a vector of that subspace.

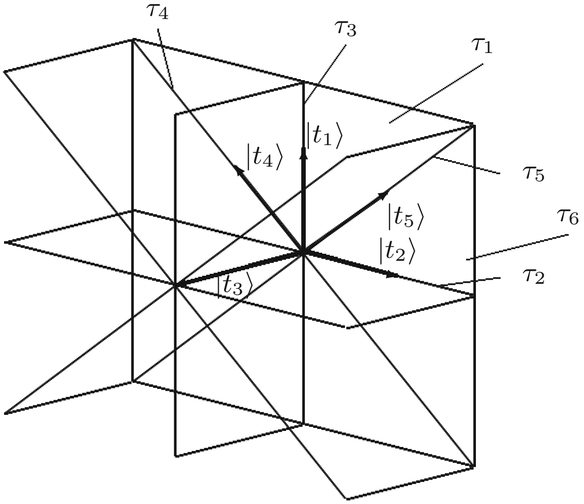


Fig. 2 A pictorial representation of features, terms, and themes. Three feature vectors $|e_1\rangle$, $|e_2\rangle$, $|e_3\rangle$ span a tridimensional vector space, but they are not depicted for the sake of clarity; the reader may suppose that the feature vectors are any triple of independent vectors. 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

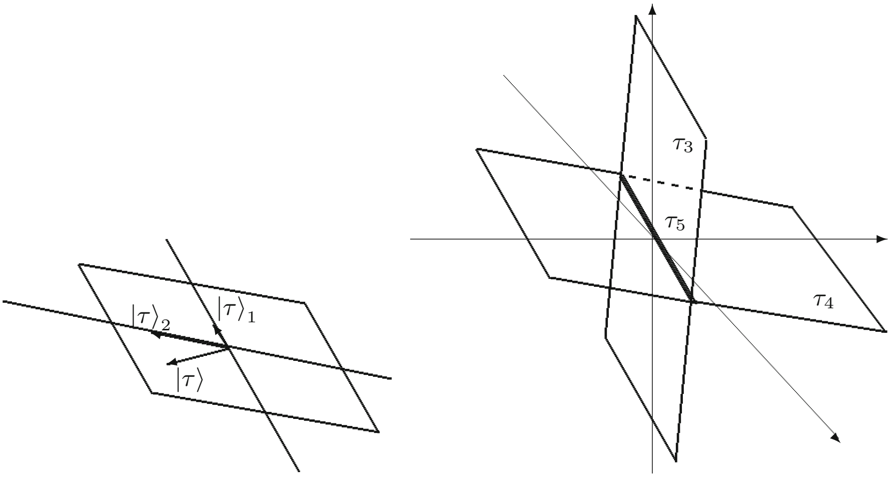
$$b_{1,1}\ |t_{1,1}\rangle + \cdots + b_{1,k_1}\ |t_{1,k_1}\rangle + \cdots + b_{n,1}\ |t_{n,1}\rangle + \cdots + b_{r,k_r}\ |t_{r,k_r}\rangle\,.$$

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



(a) (b)

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($\tau_1$, $\tau_2$)
2: for all $i = 1, 2$ do
3:     $|\tau_i\rangle \leftarrow |\tau_i\rangle \,/\, \sqrt{\langle\tau_i|\tau_i\rangle}$
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($\tau_1$, $\tau_2$)
3: $|u_3\rangle , |u_4\rangle \leftarrow$ JOIN($\tau_3$, $\tau_4$)
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

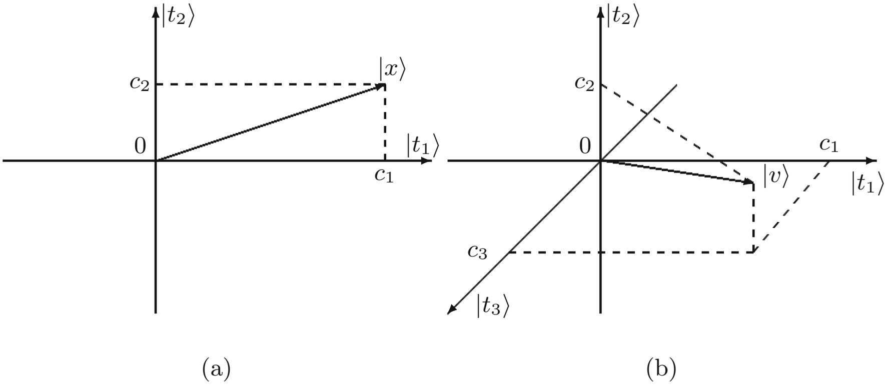


Fig. 5 How a user would build queries or documents by using the VSM

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.