Spatio-Temporal Evolution of Scientific Knowledge

Goce Trajcevski[(✉)], Xu Teng, and Shailav Taneja

Department of Electrical Engineering and Computer Science,
Northwestern University, Evanston, USA
{goce,shailavt1,xuteng2}@eecs.northwestern.edu

Abstract. In this work we take a first step towards the problem of integrating the content and the spatio-temporal aspects of the evolution of the (published) scientific knowledge. A lot of research has been invested in developing tools and search engines that will enable more efficient querying of relevant medical (and broader scientific) data from various perspectives, spanning from retrieval of similar documents/images to HCI-based flexible query-answering systems. Variety of methodologies have been developed, founded on knowledge-bases, statistics, semantic similarity, etc. and quite a few systems are available (e.g., Medline). Parallel to this, another body of research works has emerged over the past couple of decades, targeting the efficient management of mobility and spatio-temporal data. What motivates this work is the observation that fusing the data (and corresponding techniques) developed in these two broad research fields could enable novel categories of queries that can be used to investigate various evolving spatio-temporal relationships between particular scientific topics.

We present a novel model and a formalization of this confluence, in what we call *Knowledge-Evolution Trajectories* (KET). We also provide a preliminary proof-of-concept implementation that enables answering novel categories of queries pertaining to KET data with a few initial observations regarding the impact of different data-representation approaches.

1 Introduction and Motivation

Shortly after the co-emerging of the fields of spatial [1,13,23] and temporal [8] data management, the miniaturization of computing devices and advances in Global Positioning Systems (GPS) technology have spurred a plethora of applications that demanded some type of a Location-Awareness (LBS) [22]. This, originated the fields of Spatio-Temporal Databases (STDb) [9] and Moving Objects Databases (MOD) [10], addressing various aspects of managing such data – from modelling, through indexing and query processing [7,9,18,20,28]. The peculiar feature of the popular query-categories – e.g., *range, (k) Nearest-Neighbor ((k-)NN)* [26,29,30]

---

in MOD settings is that they are typically: (1) *continuous* (i.e., their answers may have to be re-evaluated based on the changes in the motion of the entities); and/or (2) *persistent* (i.e., their answers may need to be re-evaluated based both on the changes of the motion as well as the history of such changes) [7,18]. More recently, researchers have turned their attention to modelling, representing, and querying spatio-temporal trajectories which also have some kind of annotated information associated with the location and/or time – bringing about the concept of semantic (resp. symbolic, spatio-textual) trajectories [19].

Complementary to these developments, the need to reduce (or even eliminate) labor-intensive process associated with retrieval of textual documents matching particular criteria, along with the contemporary advances in information systems, have spurred the field of Information Retrieval (IR) [16,21,24] starting in the middle of the XX century. To enable effective retrieval relevant documents by various IR strategies, the documents are typically transformed into a suitable representation, often accompanied by preparation of suitable indexes [24]. In the subsequent decades, a plethora of research works[1] followed, enabling literature searches [6,17], detection of various semantic correlations among (topics) in existing publications [14], etc. In addition, several prototype systems and publicly available search engines have been generated over the years (e.g., MEDLARS [6], Medline (https://www.nlm.nih.gov/bsd/pmresources.html)).

At the heart of the motivation for this work is the observation that despite the rather long co-existence of the two fields (IR and STDb/MOD) and their respective rich histories, not much has been done in terms of exploiting the possible confluence of the two – which, as we will discuss shortly, could enable novel categories of queries of relevance for various entities, from researchers, to government agencies. For example, most of the works related to spatio-temporal aspects of medical phenomena pertain to: – modelling the temporal, spatial, and evolutionary nature of subject's conditions [4]; exploring the spreading of different chemicals, or responses/reactions to particular stimuli (e.g., [2,12].

Our goal in this work is to provide a foundation for addressing the problem of modelling the spatio-temporal aspects of the evolution of scientific knowledge and take a step towards enabling novel queries. Specifically, we aim at answering queries such as:

Q1: *Retrieve the authors and institutions located in Eastern Europe, who have published results related to the topic of heterocyclic compounds between 2005 and 2010.*
Q2: *Retrieve all the topics that were published by an institution in Pennsylvania between 2008 and 2012.*

---

[1] We do not aim at presenting a comprehensive overview of the vast body of works from the well-established field of IR in this paper (cf. [6,21]. The purpose of this section is to provide a motivation for the research presented here.

Our main contributions are two-fold:

1. We introduce the concept of *Knowledge-Evolution Trajectory* (KET) as a model to formalize the fusion between the *spatial*, *temporal*, and *content*-based aspects of scientific publications.
2. We provide a preliminary proof-of-concept implementation that demonstrates the feasibility of the proposed model. Specifically, we created a SQL Server database that: (a) contains the data pertaining to medical publications, fetched from PubMed; (b) We generated the geospatial information about each publication by using Google Map API to convert the institution name into (Lat, Lon), and then used ArcGIS to generate the values to be used by the Geometry type of SQL Server.

We also conducted some preliminary experiments which, in addition to demonstrating the feasibility of our objectives, also point out some interesting research challenges.

In the rest of this paper, Sect. 2 introduces the KET model, followed by Sect. 3 in which we describe the current case-study implementation and the experimental observations. We summarize and outline directions for future work in Sect. 4.

2   Modelling Spatio-Temporal Evolution of Scientific Literature

We now present the main aspects of the KET model. We firstly introduce the concept of symbolic trajectories.

Semantic (synonymously, Symbolic or Enriched or Spatio-textual) *Trajectories* [3,5,11,15,19] embed contextual and/or situational knowledge into location-in-time data. In a MOD [10] setting, a trajectory is modelled as a structure of the form $Tr_i = [o_{ID}, (x_{i1}, y_{i1}, t_{i1}), \ldots, (x_{ik}, y_{ik}, t_{ik})]$, where x_{ij} and y_{ij} $(1 \leq j \leq k)$ are the coordinates of the location $(l_{ij} = (x_{ij}, y_{ij}))$ of the object with a unique oID, obtained at time instant t_{ij}. In-between two consecutive updates, the object's motion is approximated in accordance with some kind of an interpolation. STs attempt also to describe the kinds of activities associated with a particular location and time – e.g., "stop", "move", "walk", "eat", etc. Formally (cf. [5,19]), a semantic trajectory ST_i is a sequence of so-called, semantic episodes $se_{i,m}$ as follows:

$ST_i = [se_{i1}, se_{i2}, se_{i3}, \ldots se_{im}]$, and each semantic episode is a tuple of the form:

$$se_{ij} = (da_{ij}, sp_{ij}, x_{ij}^{in}, y_{ij}^{in}, t_{ij}^{in}, x_{ij}^{out}, y_{ij}^{out}, t_{ij}^{out}, tagList_{ij})$$

where:

– da_{ij} = defining annotation; typically expressing an activity (verb) such as stop or move.
– sp_{ij} = semantic location/position of the activity, like a POI (e.g., a museum, restaurant, zoo), home, work, etc.

- t_{ij}^{in} and t_{ij}^{out} = entry/exit times of a semantic position.
- $x_{ij}^{in}, y_{ij}^{in}, x_{ij}^{out}, y_{ij}^{out}$ = entry/exit coordinates of a semantic position.
- tagList$_{ij}$ = any additional semantic information, like transportation mode, additional activity description (e.g., eat), etc.

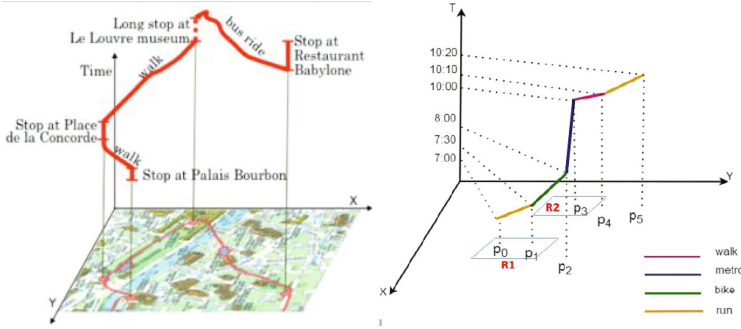constitute the j-th semantic episode of the i-th semantic trajectory.

Fig. 1. Semantic trajectory and spatio-temporal range querying

Left portion of Fig. 1 illustrates the concept of a semantic trajectory (cf. [19]), and the right portion (cf. [15]) illustrates yet another way of visualizing a semantic trajectory (i.e., color-coded activity) along with the semantics of processing the query *Retrieve all the individuals who were running in the region R1, between 7:00 and 8:00 AM.*

There are two fundamental observations when it comes to the existing model of semantic trajectories, and the KET model that we are proposing:

1. There is no concept of a "motion", as commonly perceived in MOD trajectories (even when augmented with annotation). The scientific publications are associated with spatial data – e.g., the locations of the institution of the participating authors, and those are not mobile entities. However, there is an evolution over the temporal dimension that, in part, is associated with spatial values.
2. The scientific publications have a lot more contextual attributes, and those attributes are composite/richer. Namely, considering a typical meta-data[2], they contain:

- *Title*;
- *Category* (in accordance with an established nomenclature of the corresponding field; possibly a *set* of such categories, augmented with a set of *keywords*).
- A *set* of *authors names* and the corresponding *institution* of his affiliation.

---

[2] Aside from the main body of the text of the respective publications, or other attributes associated with, e.g., publishers, forum/venue, etc...

It is precisely the elements of the *institution* that contain an *implicit spatial value*, which we use for enabling the novel categories of queries.

Based on **(1)** and **(2)** above, we define the concept of *Scientific Publication Point* (SPP) as follows:

$SPP_i =$
$(P_{ID}, Title, category, [(author_{1i}, inst_{1i}, x_{1i}, y_{1i}), \ldots, (author_{ki}, inst_{ki}, x_{ki}, y_{ki})], T_{pub})$

We re-iterate that, in practice, most of the bibliographical data sources contain the name of the institution (along with its postal-address) for each author – however, not the actual coordinates for the corresponding address. This, in turn, eliminates the possibility of asking any queries containing predicates ranging over spatial domains. However, such queries may provide insight into data/trends that could influence both government funding as well as institutional/individual collaboration plans, as exemplified by the query: *Retrieve all the institutions within 100 Km from the coastal line in China, which have received more than 10M renminbi research funding in the last 4 years for medical research, but have published less than 20 articles on the topic of cytostatics.*
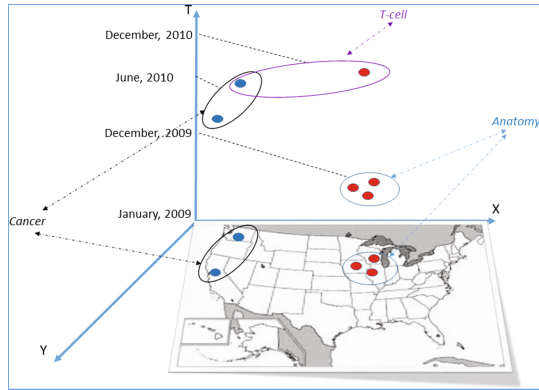

Fig. 2. KET for Geo-constrained Query

Assuming that the temporal value in each SPP_i (i.e., $SPP_i.T_{pub}$) is represented at a uniform particular level of granularity (e.g., (*month, year*)), we now present formally the model for a KET (*Knowledge-Evolution Trajectory*) defined as follows:

Definition 1. *A Knowledge-Evolution Trajectory is a sequence*
  $[\alpha_1(SPP^{(1)}), \alpha_2(SPP^{(2)}), \ldots, \alpha_n(SPP^{(n)})]$ *where:*

– α_i *denotes an operator from relational algebra or a spatial predicate, applied to SPP_i*
– *For any pair $SPP^{(i)}$, $SPP^{(j)}$, $(i < j) \Rightarrow SPP^{(i)}.T_{pub} < SPP^{(j)}.T_{pub}$*

Given the definition of SPP_i, the role of α_i in Definition 1 is to extract the proper content of interest, the evolution of which needs to be queried. Thus, for example, we can focus on a particular author by applying $(SPP_i).author = 'Jones'$. However, the main benefit of the proposed model is that we can also apply spatial predicates such as: $(SPP_i).(x_{mi}, y_{mi})IN'Pennsylvania'$.

Clearly, the KET model generalizes the traditional model of a trajectory used in MOD literature and, for that matter, also generalizes the semantic trajectories.

Figure 2 shows an example of a KET corresponding to an answer of the query *Retrieve all the collaborative works between Midwest-based and California-based institutions, between January 2009 and December 2010.* As can be seen, instead of a traditional (x, y) point, we now have collection of Geo-points that constitute each one of the trajectories of the answer. Moreover, we can also see that a particular Geo-point can participate in > 1 KET – for as long as the constraints of the query are satisfied. Thus, we have a collaboration between an institution from California and Illinois on a publication related to T-cell studies, in December of 2010.
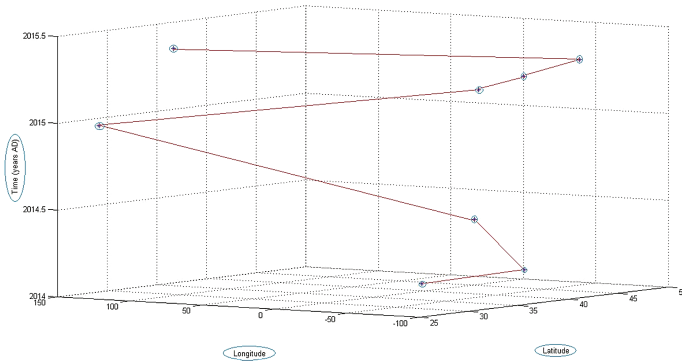

Fig. 3. KET for participation constraint query

Figure 3 illustrates another example of a KET – which visually (and type-wise) has a highest resemblance of a traditional moving point-object trajectory. However, it shows an example of an actual answer from our implementation, corresponding to the query *Retrieve all the institutions that have published an article on cytostatics in which all the authors were from a same institution, between January, 2014 and July 2015.*

We close this section with another example-query the answer of which, in some sense, does resemble spatio-temporal trajectories, but yet it has its own distinct semantics of the temporal evolution.

Figure 4 shows the answer to the query *Retrieve which topics/categories had most publications, for the researchers from the SouthWest University, as well as for the ones from Arizona State University, between 2008 and 2012.*
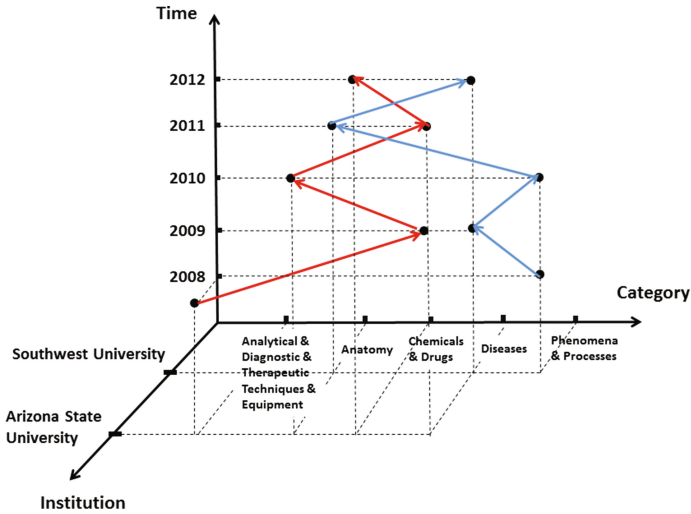
Fig. 4. Following most popular topics per institution

3   Case-Study: Spatio-Temporal Evolution in PubMed

We now describe in detail the current implementation of our proof-of-concept system for which the context is restricted to the PubMed data, pertaining to the various medical publications available on MedLine. We used SQL Server 2014 and we wrote python scripts to extract the data from PubMed and populate the tables[3]. As mentioned earlier, the PubMed data does not contain geo-spatial information[4] – whereas SQL Server provides two different geospatial data types: `Geometry` type and `Geography` type. Thus, in our prototype system, we selected [Publish ID], [Title], [Author] (including their names and institutions), [Publish Date] from the returned XML records from PubMed – however, in addition, we populated the entries for the corresponding `Geometry` type in which the coordinate system is World Geodetic System (WGS) 1984. To populate the corresponding values for the geospatial information, we used a two-step procedure:

(1) Google Map API was used to convert the institution name into (Lat, Lon) pair of values.
(2) Subsequently, ArcGIS was used to convert from (float Lat, float Lon) into the corresponding `Geometry` type.

---

[3] We note that all the data, code for the queries, as well as the scripts used to generate the values for the spatial attributes, is publicly available at https://github.com/ ShailavTaneja/PubMedDerivedDataAndImplementation.

[4] The description of the standard meta-data used in PubMed is available at https://www.ncbi.nlm.nih.gov/books/NBK3828/#publisherhelp.Example_of_a_ Standard_XML.

This is what enabled us to specify queries that can capture the spatio-temporal evolution of the knowledge represented in scientific – in this case, medical – publications.

In the first iteration, we had a "naïve" representation of the data residing in a single table, with attributes:

```
- [Publish ID] varchar(50),
- [Title] varchar(500),
- [Publish Date] date,
- [Author Name] varchar(50),
- [Institution Name] varchar(150),
- [Geo Information] geometry,
- [Subcategory] varchar(150),
- [Category] varchar(150)
```

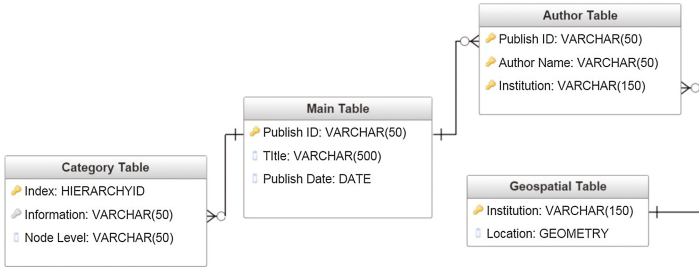The table had 409702 rows in total.

Fig. 5. Normalized database schema

Subsequently, to eliminate the redundancy, we normalized the naïve table, and the schemata that we used in the implementation is shown in Fig. 5, corresponding to the following tables:

– **Main** – with three attributes: [Publish ID] varchar(50), [Title] varchar(500), [Publish Date] date. 32768 row in total
– **Author** – with three attributes: [Publish ID] varchar(50), [Author Name] varchar(50), [Institution] varchar(150). 188786 rows in total.
– **Category** is a table capturing the hierarchy of the categories. It has three attributes: [Index] hierarchyid, [Information] varchar(150), [Node Level] varchar(50). 68652 rows in total. **Hierarchyid** is a special data type and works as the key of Category table. / represents root, /*/ represents the first level and so on
– **Geospatial** – with two attributes: [Institution] varchar(150), [Location] geometry. 1316 rows in total.

In the sequel, we describe an example of the difference between evaluating queries in the naïve representation and the normalized one. Consider the following:

$\mathbf{Q}_{topic}$: *Retrieve the* KET *for publications addressing the topic of diagnosis, in the period of October 1, 2012 – September 1, 2016.*

The query returns a set of temporally annotated points (geo locations) for the topic, during that period.

The corresponding SQL statement for the naïve implementation is:

```
SELECT DISTINCT [Publish ID], [Publish Date], Location.STX, Location.STY
FROM NativeTable
WHERE ([Publish Date] between '2012-10-01' and '2016-09-01') and
      ([Subcategory] = 'diagnosis')
ORDER BY [Publish Date]
```

The SQL implementation for the same query in the normalized version is:

```
DECLARE @catLevel hierarchyid;
SET @catLevel = (SELECT Index
                        FROM Category
                        WHERE Information = 'diagnosis')

SELECT DISTINCT MainTable.[Publish ID], MainTable.[Publish Date],
                                Location.STX, Location.STY
FROM MainTable, Author, Geospatial
WHERE (MainTable.[Publish Date] between '2012-10-01' and '2016-09-01') and
(MainTable.[Publish ID] in
 (SELECT Information
    FROM Category
    WHERE Index.IsDescendantOf(@catLevel) = 1 and [Node Level] = 'Publish ID'))
            and
      (MainTable.[Publish ID] = Author.[Publish ID]) and
      (Author.Institution = Geospatial.Institution)
ORDER BY MainTable.[Publish Date]
```

To illustrate the impact of the different database representation, we first illustrate the benefits in terms of eliminating the redundancy via the normalized representation:

As shown in Fig. 6, the naïve implementation requires approximately five times more space than the normalized one.

When it comes to the efficiency of the execution with each implementation we report the corresponding measurements observed when executing $\mathbf{Q}_{topic}$ (labeled **Q1** in Fig. 7) and the query

Q2: *Retrieve the works jointly co-authored by Masaki Mori and Yuichiro Doki, between 2009–2010 and 2012–2013*

 on a Windows 10 machine, with Intel(R) Core(TM) i5-5257U CPU (2.70 GHz, 2701 Mhz), with 2 Cores (4 Logical Processors) and 8 GB of RAM.

As shown in Fig. 7, the normalized implementation also yields a much more efficient execution than the naïve one.
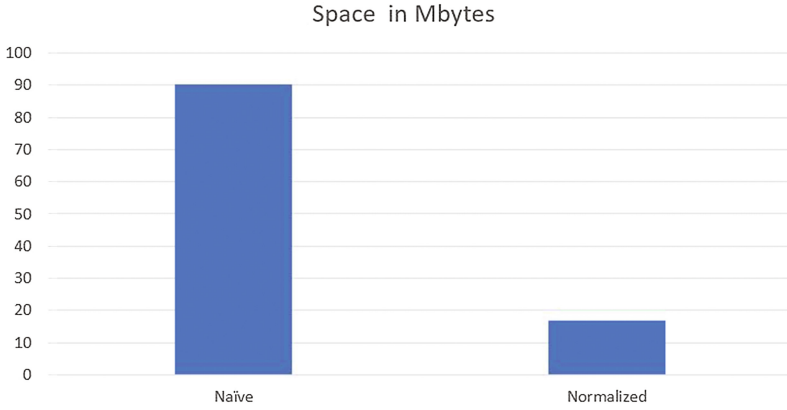
Space in Mbytes

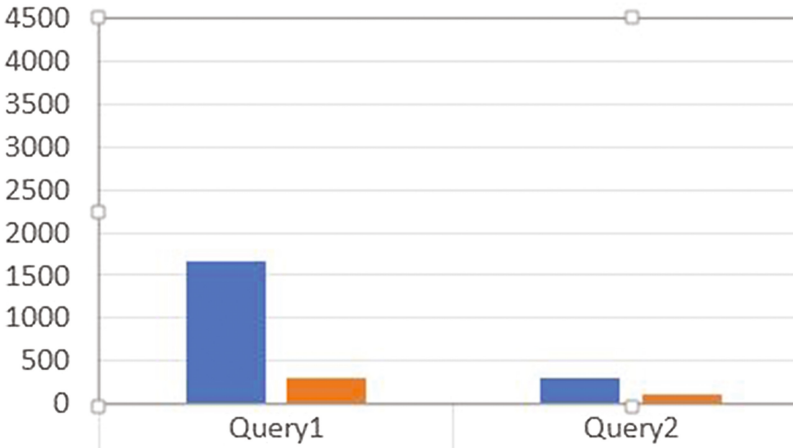**Fig. 6.** Space requirements

Fig. 7. Efficiency of execution (in milliseconds)

4   Conclusions and Future Work

We addressed the problem of modelling and querying the spatio-temporal evolution of the knowledge recorded in scientific publications, and took the first steps towards providing a formalism to capture such evolution across the (geo)spatial domain, as well as other contextual attributes. We proposed the concept of KET (Knowledge-Evolving Trajectories) as a possible unification between two fields (Information Retrieval and Spatio-Temporal/Moving Objects Databases). This type of unification provides opportunities for novel query categories which, to our knowledge, have not been formally treated in the literature. In addition, we provided a proof-of-concept implementation of our approach for the data available in PubMed, and compared the impacts of a naïve design of the database vs. a normalized one – albeit for limited set of queries. Our initial evaluations have

demonstrated that in addition to reducing the space requirements, the normalized approach also enables faster execution of the KET-based queries.

Related Works: There are plethora of works in each of the fields of IR [6, 14, 17, 21] and STDb/MOD [7, 9, 10, 18, 20, 23, 28] – to mention but a few. The novelty of our proposed approach is to provide a formalism for fusing these works, along with the integration of the respective existing datasets, enabling novel query-categories. The closest formalism to our proposed approach is the one of semantic/symbolic trajectories [19]. However, as we argued, this approach: (1) has too simplistic model of the spatio-temporal evolution; and (2) is lacking the - dimensionality of the contexts typically associated with scientific publications.

We believe to have scratched the surface of a direction that may be of interest in many applications of societal relevance and, moreover, can pose interesting challenges. As part of our future work, we are planning to extend the KET model, and augment the current implementation so that it can incorporate different publications' data sources. In addition, although we have provided some preliminary discussions related to the efficiency, a challenging problem is to address the efficient processing different types of KET-based queries.

Last, but not the least, it seems rather intuitive that investigations along the direction of warehousing spatio-temporal evolution of scientific publications along with further semantic similarity searches [25, 27] data may yield novel categories of analytical queries.

References

1. Bedard, Y., Merrett, T., Han, J.: Fundamentals of spatial data warehousing for geographic knowledge discovery. Geogr. Data Min. Knowl. Discov. **2**, 53–73 (2001). Taylor and Francis
2. Bilgen, M., Abbe, R., Liu, S.J., Narayana, P.A.: Spatial and temporal evolution of hemorrhage in the hyperacute phase of experimental spinal cord injury: in vivo magnetic resonance imaging. Magn. Ressonance Med. **43**(4), 594–600 (2000)
3. Bogorny, V., Renso, C., de Aquino, A.R., de Lucca Siqueira, F.: Constant - A conceptual data model for semantic trajectories of moving objects. GIS **18**(1), 66–88 (2014)
4. Chu, W.W., Cardenas, A.F., Taira, R.T.: Kmed: A knowledge-based multimedia medical distributed database system. Inf. Sci. **20**(2), 75–96 (1995)
5. Damiani, M.L., Güting, R.H.: Semantic trajectories and beyond. In: Proceedings of IEEE - MDM, pp. 1–3. Brisbane, Australia (2014)
6. Dee, C.R.: The development of the medical literature analysis and retrieval system (medlars). J. Med. Library Assoc. **94**(5), 416–425 (2007)
7. Ding, H., Trajcevski, G., Scheuermann, P.: Towards efficient maintenance of continuous queries for trajcectories. GeoInformatica **12**(3), 255–288 (2008)
8. Etzion, O., Jajodia, S., Sripada, S. (eds.): Temporal Databases: Research and Practice. LNCS, vol. 1399. Springer, Heidelberg (1998)
9. Güting, R.H., Böhlen, M.H., Erwig, M., Jensen, C.S., Lorentzos, N., Schneider, M., Vazirgiannis, M.: A foundation for representing and queirying moving objects. ACM TODS **25**, 1–42 (2000)
10. Güting, R.H., Schneider, M.: Moving Objects Databases. Morgan Kaufmann, San Francisco (2005)

11. Güting, R.H., Valdés, F., Damiani, M.L.: Symbolic trajectories. ACM Trans. Spat. Algorithms Syst. **1**(2), 7:1–7:51 (2015)
12. Hirano, Y., Stefanovic, B., Silva, A.C.: Spatiotemporal evolution of the fmri response to ultrashort stimuli. J. Neurosci. **31**(4), 1440–1447 (2011)
13. Hjaltason, G.R., Samet, H.: Distance browsing in spatial databases. ACM Trans. Database Syst. **24**(2), 265–318 (1999)
14. Hristovski, D., Kastrin, A., Dinevski, D., Rindflesch, T.C.: Constructing a graph database for semantic literature-based discovery. In: MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, p. 1094 (2015)
15. Issa, H.: Spatio-textual trajectories: models and applications. PhD thesis, Universita degli studi di Milano (2017)
16. Korfhage, R.: The impact of personal computers on library-based information systems. SIGIR Forum **12**(4), 10–13 (1978)
17. Lowe, H.J., Barnett, G.O.: Understanding and using the medical subject heading (mesh) vocabulary to perform literature searches. J. Am. Med. Assoc. **271**(14), 1103–1108 (1994)
18. Mokbel, M.F., Aref, W.G.: SOLE: scalable on-line execution of continuous queries on spatio-temporal data streams. VLDB J. **17**(5), 971–995 (2008)
19. Parent, C., Spaccapietra, S., Renso, C., Andrienko, G.L., Andrienko, N.V., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A., de Macêdo, J., Pelekis, N., Theodoridis, Y., Yan, Z.: Semantic trajectories modeling and analysis. ACM Comput. Surv. **45**(4), 42 (2013)
20. Pelanis, M., Saltenis, S., Jensen, C.S.: Indexing the past, present, and anticipated future positions of moving objects. ACM Trans. Database Syst. **31**(1), 255–298 (2006)
21. Salton, G.: Automatic Text Processing. Addison Wesley, Massachusetts (1989)
22. Schiller, J.H., Voisard, A. (eds.): Location-Based Services. Morgan Kaufmann, San Francisco (2004)
23. Shekhar, S., Chawla, S.: Spatial Databases: A Tour. Prentice Hall, New Jersy (2003)
24. Taine, S.I.: New program for indexing at the national library of medicine. Bull. Med. Libr. Assoc. **47**(2), 117 (1959)
25. Trajcevski, G., Donevska, I., Vaisman, A.A., Avci, B., Zhang, T., Tian, D.: Semantics-aware warehousing of symbolic trajectories. In: Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS 2015, pp. 1–8, 3–6 November 2015, Bellevue, WA, USA (2015)
26. Trajcevski, G., Tamassia, R., Cruz, I., Scheuermann, P., Hartglass, D., Zamierowski, C.: Ranking continuous nearest neighbors for uncertain trajectories. VLDB J. **20**(5), 767–791 (2011)
27. Vaisman, A.A., Zimányi, E.: Data Warehouse Systems: Design and Implementation. Data-Centric Systems and Applications. Springer, Heidelberg (2014)
28. Xing, X., Mokbel, M.F., Aref, W.G., Hambrusch, S.E., Prabhakar, S.: Scalable spatio-temporal continuous query processing for location-aware services. In: International Conference on Scientific and Statistical Database Management (SSDBM) (2004)
29. Xiong, X., Mokbel, M.F., Aref, W.G.: Sea-cnn: Scalable processing of continuous k-nearest neighbor queries in spatio-temporal databases. In: ICDE, pp. 643–654 (2005)
30. Yu, X., Pu, K.Q., Koudas, N.: Monitoring k-nearest neighbor queries over moving objects. In: ICDE, pp. 631–642 (2005)