Data & Knowledge Engineering 67 (2008) 234–259

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

The thematic and citation landscape of *Data and Knowledge Engineering* (1985–2007) ☆

Chaomei Chen [a], Il-Yeol Song [a,*], Xiaojun Yuan [b], Jian Zhang [a]

[a] College of Information Science and Technology, Drexel University, United States
[b] College of Computing and Information, University at Albany, State University of New York, United States

ARTICLE INFO

Article history:
Available online 28 May 2008

Keywords:
Structural and temporal patterns
Domain analysis
Scientometrics
CiteSpace
Thematic analysis
DKE

ABSTRACT

The thematic and citation structures of *Data and Knowledge Engineering* (DKE) (1985–2007) are identified based on text analysis and citation analysis of the bibliographic records of full papers published in the journal. Temporal patterns are identified by detecting abrupt increases of frequencies of noun phrases extracted from titles and abstracts of DKE papers over time. Conceptual structures of the subject domain are identified by clustering analysis. Concept maps and network visualizations are presented to illustrate salient patterns and emerging thematic trends. A variety of statistics are reported to highlight key contributors and DKE papers that have made profound impacts.

1. Introduction

Data and Knowledge Engineering (DKE) is a premier journal focusing on the common research themes in the areas of database systems and knowledge base systems. The major aim of the journal is to identify, investigate and analyze the underlying principles in the design and effective use of these systems. DKE publishes original research results, technical advances, and new experiments concerning data engineering, knowledge engineering, and the interface of these two fields. Under the leadership of Professors Peter Chen and Reind van de Riet, DKE has served its roles as a leading scholarly journal.

Since its first publication in 1985, DKE has reached a world-wide audience of researchers, designers, managers and users. DKE also has covered a wide spectrum of research topics such as representation and manipulation of data and knowledge, architectures of database and knowledge-based systems, construction of data and knowledge bases, tools and methodologies for developing data and knowledge bases, applications, case studies, and management issues. DKE has made significant contributions and impacts on the advances of data and knowledge engineering.

In order to celebrate the 25th anniversary of DKE, we embarked an extensive analysis of papers that have been published in DKE over the last 24 years. Our analysis includes authorship analysis, coauthorship analysis, thematic trends and conceptual structures, visualization of topic maps and various citation networks. The analysis is based on bibliographic records and citation information retrieved from the Web of Science (WoS), Scopus (http://www.scopus.com/), and ScienceDirect (http://www.sciencedirect.com/).

☆ CiteSpace is available at http://cluster.cis.drexel.edu/~cchen/citespace. Color versions of the figures are available at http://cluster.cis.drexel.edu/~cchen/papers/2008/dke/figures.

* Corresponding author.
E-mail addresses: chaomei.chen@cis.drexel.edu (C. Chen), songiy@drexel.edu (I.-Y. Song), xy454777@albany.edu (X. Yuan), jz85@drexel.edu (J. Zhang).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.05.004

2. A longitudinal study of DKE

Our study primarily utilizes CiteSpace[1] as an analytic and visualization tool. CiteSpace is a freely available Java application for analyzing and visualizing emerging trends and citation patterns in scientific literature [1,2]. CiteSpace is continuously evolving to incorporate a variety of visual analytic functions. CiteSpace is designed to simplify the analysis of scientific literature by enabling users to find salient patterns from a diverse range of visual attributes. CiteSpace follows a simple model of the dynamics of scholarly communication in which a transient body of scientific papers, collectively known as a research front, makes reference to a group of papers in the literature, which is called the intellectual basis [1]. Based on this model, CiteSpace aims to make it easy for users to identify some special classes of papers in terms of *landmarks* by citation popularity, *hotspots* by abrupt increases of citations they received, and *pivotal papers* that are strategically positioned in co-citation networks. Landmark papers are depicted by their large-sized tree-ring circles. Hotspots are shown as nodes with a red rim. Pivotal papers are shown as nodes with a purple rim.

The general procedure of analysis and visualization with CiteSpace is outlined as follows. Technical details are provided elsewhere [1,2].

1. Identify a knowledge domain. In this study, the knowledge domain is defined by DKE papers and their citations.
2. Data collection. We collect bibliographic records and citation data associated with DKE papers from three sources: ScienceDirect,[2] Scopus,[3] and the Web of Science.[4] See Section 3.1 for detail.
3. Extract noun phrase terms from titles, abstracts, descriptors, and identifiers of citing articles in the dataset. The burstness of the extracted terms is detected for abruptly increased frequencies of specific terms [3]. Burst terms are used to capture fast-growing interests.
4. Time slicing. Specify the range of the entire time interval and the length of a single time slice.
5. Threshold selection. CiteSpace allows users to specify three sets of threshold levels for citation counts, co-citation counts, and co-citation coefficients. Citation counts are the number of times a publication is cited by DKE papers. Two publications are called co-cited if a paper cites both of them. Co-citation counts for a given pair of publications are the number of papers in our dataset that cite the pair. Co-citation coefficients are normalized co-citation counts over each time slice. The specified thresholds are applied to three time slices, namely, the earliest slice, the middle one, and the last one. Linear interpolated thresholds are assigned to the rest of slices. CiteSpace supports networks of four types of nodes and three types of links. Nodes include authors, papers, journals, and burst terms, whereas links may represent co-occurrence, co-citation, or referential links.
6. Pruning and merging. Pathfinder network scaling [4,5] is the default option in CiteSpace for network pruning [2,4]. Users choose whether or not to apply the scaling operation to individual networks. CiteSpace merges individual networks by taking a set union of all the vertices and selecting links that do not violate a triangle inequality condition in overlapping areas between networks. Users can choose whether or not to prune the merged network as a whole.
7. Layout. CiteSpace supports a standard graph view and a time-zone view.
8. Visual inspection. CiteSpace enables users to interact with the visualization of a knowledge domain in several ways. The user may control the display of visual attributes and labels as well as a variety of parameters used by the underlying layout algorithms.
9. Verify pivotal points. The significance of a marked pivotal point can be verified by asking domain experts, for example, the authors of pivotal-point articles, and/or examining the literature, such as passages containing citations of a pivotal-point article. A particularly interesting direction of research is the development of tools that can automatically summarize the value of a pivotal point. Digital libraries, automated text summarization, machine learning, and several other fields are among the most promising sources of input.

3. Methods

3.1. Data collection

We searched for bibliographic records of the journal *Data & Knowledge Engineering* (DKE) in three sources, namely ScienceDirect, Scopus, and the Web of Science. ScienceDirect has the most extensive coverage of papers published in DKE. We found 993 bibliographic records of DKE papers in ScienceDirect, 865 DKE records in Scopus, and 774 records in the Web of Science.

The three sources vary in their accessibility of cited references. Although users can access cited references in interactive modes, the Web of Science provides the most reliable download function. The Web of Science imposes a 500-record limit per download request. Multiple download operations are necessary to retrieve a dataset with more than 500 records. Scopus

[1] http://cluster.cis.drexel.edu/~cchen/citespace/.
[2] http://www.sciencedirect.com/.
[3] http://www.scopus.com/.
[4] http://scientific.thomson.com/products/wos/.

provides the second most reliable function for downloading cited references. Scopus imposes a 2000-record limit per download request. ScienceDirect does not provide functions for downloading cited references. Table 1 summarizes the records retrieved from the three sources.

Fig. 1 shows the distribution of bibliographic records from the three sources. Since the set of 993 bibliographic records from ScienceDirect is the most comprehensive, the ScienceDirect set is used for subsequent text analyses, including the analysis of thematic trends and clustering analysis. The Web of Science (WoS) set has the most comprehensive coverage of cited references since 1994. The Scopus set is the only citation dataset that covers the period of 1985–1993. It is almost identical to the WoS set from 1994 till 2000, but misses a considerable number of records between 2001 and 2006. Therefore, the WoS set is used as the primary source for citation analysis and supplemented with the Scopus set.

3.2. Analysis of authorship and coauthorship

The analysis of authorship focuses on the productivity of DKE authors and their impact in terms of citations they received. The analysis aims to provide a useful glimpse of the dynamic structure of the contributing research community, which is typically perceived as an invisible college because such insights are often privileged knowledge to insiders, i.e. experienced domain experts [6]. A revelation of this type will be particularly useful for new comers to the research community and for anyone who is looking for potential collaborators, especially for interdisciplinary research, or looking for reviewers or panelists. The analysis of coauthorship is also valuable in this vein [7]. Besides the practical reasons, the study of authorship and coauthorship is an important approach to the understanding of scholarly communication and knowledge diffusion.

The productivity analysis is primarily based on the ScienceDirect dataset with references to the Scopus and the WoS sets, whereas the impact analysis is based on the WoS dataset. The productivity of DKE authors is measured by the number of DKE papers one has published as a co-author as well as the first author. The productivity analysis also identifies the most productive institutions based on the number of DKE papers they published.

The impact analysis identifies the most influential DKE authors based on the number of citations attributed to their names in the WoS dataset. In addition, author co-citation analysis is included in our study to identify higher-order connectivity patterns between authors. Author co-citation analysis identifies both DKE authors and none-DKE authors. A none-DKE author in this context refers to a researcher who has never published any DKE paper, but has published papers that have been cited by DKE papers.

Coauthorship is defined between a pair of researchers if they are co-authors of at least one DKE paper. DKE coauthorship is analyzed in two ways: one as a coauthorship network and the other as a geospatial overlay on a world map. A coauthorship network is generated by CiteSpace based on the ScienceDirect dataset (1985–2007). The coauthorship network is used to identify key players in the context of DKE in terms of their betweenness centrality scores. In a network, the betweenness centrality of a node measures the extent to which the node plays a role in pulling the rest of nodes in the network together. The higher the centrality of a node, the more strategically important the node is. The geospatial overlay is also generated by CiteSpace. It is viewable with Google Earth. The geospatial overlay is useful for identifying collaborative patterns in association with geographical proximity.

3.3. Analysis of thematic trends and conceptual structures

Scientific literature contains both persistent and transient elements [8]. The transient aspect of scientific literature can be characterized by corresponding thematic trends, whereas the persistent aspect can be characterized by salient conceptual structures. The analysis of thematic trends is based on the concept of burst detection [3]. Salient conceptual structures can be identified through clustering analysis. Identified thematic trends and conceptual structures can improve our understanding of what topics are hot, how long a particular thematic trend is expected to grow, and how a variety of topics fit on a global intellectual picture.

A burst detection algorithm is typically applied to a frequency function $F(t)$ defined over a time interval T and finds subintervals in which $F(t)$ is elevated statistically with reference to the dataset as a whole. In our study, burst detection algorithms are applied to terms found in the abstracts of DKE papers in the ScienceDirect dataset. Later on, we also apply burst detection algorithms to citation frequencies of papers cited by the WoS dataset.

Table 1
Summary of the search results

		ScienceDirect	Scopus	Web of Science
Bibliographic records	Dates	1985–2008	1985–2008	1994–2008
Text analysis	# Items	**993**	865	774
Cited references	Dates	N/A	1996–2008	1994–2008
Citation analysis	# Items	0	9143	**22,574**

Document type = article, subject area = computer science.

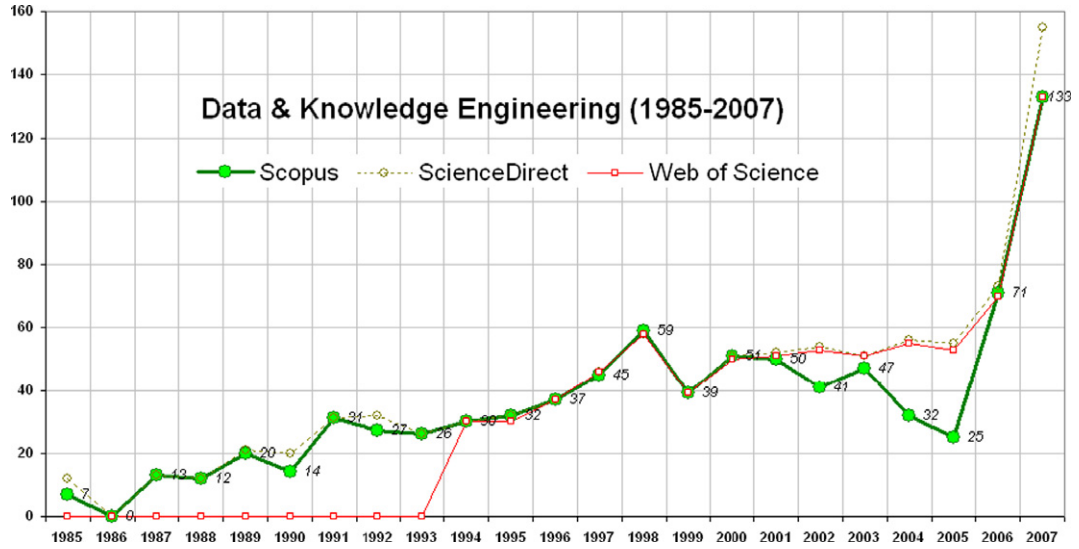


Fig. 1. Distributions of bibliographic records of DKE papers.

In our study, terms refer to noun phrases extracted from the title and abstract of DKE papers. In this paper, noun phrase extraction is limited to noun phrases consisting of 2–4 words such as *conceptual modeling* and *deductive database*. If it is detected that the term *conceptual modeling* experienced a period of burst between 1997 and 2000, then it means that the topic of *conceptual modeling* was extraordinarily active during this period as far as the forum sustained by DKE is concerned. Therefore, we use such periods of burst to characterize thematic trends manifested through DKE papers.

In addition to the burstness of noun phrases extracted from titles and abstracts of DKE papers, the burstness of keywords assigned to DKE papers is also taken into account in our study. In general, keywords identify thematic topics at a higher conceptual level than noun phrases. For example, noun phrase *clustering algorithms* may be associated with keyword *data mining*. Thus, the burstness of keywords offers additional macroscopic insights.

Interrelationships between noun phrases at microscopic levels and keywords at macroscopic levels are represented in the form of concept maps of both types of entities and their relations. The association between a noun phrase and a keyword is established based on how frequently they are found in the same bibliographic record. The more frequently they appear together, the stronger the association is. Noun phrase extraction and conceptual mapping of keyword-term relationships are both generated by CiteSpace.

The aim of clustering analysis is to identify salient conceptual structures. The salient conceptual structure in our case is derived from word occurrences in the DKE abstracts. We use probabilistic latent semantic analysis (pLSA) to identify salient topic clusters. pLSA uses a probabilistic latent variable model to associate the latent class variables ($\boldsymbol{z}$) with observable co-occurrence variables of words ($\boldsymbol{w}$) and documents ($\boldsymbol{d}$) [9]. Each cluster can be represented by words that are highly expected given the latent variable, i.e. by $p(\boldsymbol{w}|\boldsymbol{z})$. pLSA is a special case of non-negative matrix factorization (NMF), which approximates a matrix by multiplications of non-negative matrices. The Lemur toolkit[5] is used for the clustering analysis.

3.4. Analysis of citation patterns

The third type of units of analysis is DKE papers. Scientific publications can be seen as a proxy of knowledge elements. We first identify the most cited DKE papers so as to underline the scope and depth of DKE as a forum for scholarly communication. The first question to be addressed is which DKE papers are highly cited in the Web of Science as a whole as well as by the WoS dataset in particular. It is quite possible that a DKE paper is cited not only by other DKE papers, but also by papers published elsewhere. Each WoS record comes with a times cited (TC) field, which indicates the number of times that the paper is cited by the Web of Science as a whole. In other words, the value of TC is often higher than the number of citations made by papers in the specific dataset retrieved from the Web of Science, i.e. in this case, the WoS dataset of DKE.

In this part of the study, we also examine the burstness of citations to papers cited by DKE papers. Cited references with burst highlight the focus of DKE's contributors and how the focus changes over time. Cross-referencing between burst terms and burst references can lead to further insights into the underlying thematic trends.

[5] http://www.lemurproject.org/.

Table 2
Authors of more than five DKE papers from three sources

Source	WOS (1994–3/20/08) [735]		Scopus (1998–3/20/08) [842]		ScienceDirect (1995–2007) 932	
Rank	# Papers	Authors	# Papers	Authors	# Papers	Authors
1.	16	Tan, K.L.	16	Tan, K.L.	15	Tan, K.L.
2.	12	Bhowmick, S.S.	13	Bhowmick, S.S.	14	Bhowmick, S.S.
3.	11	Manolopoulos, Y.	12	Manolopoulos, Y.	13	Manolopoulos, Y.
4.	10	Li, W.S.	10	Li, W.S.	12	van de Riet, R.P.
5.	9	Madria, S.	10	Hunter, A.	11	Storey, V.C.
6.	9	Hunter, A.	10	Storey, V.C.	10	Hunter, A.
7.	8	Lee, S.	8	Madria, S.	9	Madria, S.
8.	7	Kim, H.J.	8	Bertino, E.	9	Li, W.S.
9.	7	Ooi, B.C.	8	Proper, H.A.	9	Proper, H.A.
10.	7	Storey, V.C.	6	Ooi, B.C.	9	Bertino, E.
11.	6	van der Aalst, W.M.P.	6	Orlowska, M.E.	8	Lee, S.
12.	6	Rundensteiner, E.A.	6	Saake, G.	8	Orlowska, M.E.
13.	6	Ursino, D.	6	van de Riet, R.P.	7	Kim, H.J.
14.	6	Orlowska, M.E.	6	Sellis, T.	7	Rundensteiner, E.A.
15.	6	Bertino, E.			7	van der Weide, T.P.
16.	6	Thalheim, B.			7	Saake, G.
17.	6	Proper, H.A.			7	Sellis, T.
18.					6	Thalheim, B.
19.					6	Paton, N.W.
20.					6	Bell, D.A.

Finally, document co-citation analysis (DCA) is included to analyze the global connectivity patterns of papers cited by the DKE papers. The citation image and the concept maps of keywords and noun phrases form a rich picture of the role of DKE in data and knowledge engineering. Results in this part of the analysis are obtained through CiteSpace.

4. Results

Results are organized in the same structure as Section 3. First, we report the results regarding DKE authorship and coauthorship. Second, we describe the results of thematic trends and conceptual structures. Third, we summarize the results of citation patterns found in the DKE dataset.

4.1. Authorship and coauthorship

4.1.1. Authorship

Table 2 lists the most productive DKE authors who have published more than five DKE papers. The list consists of paper counts from all three data sources. As we discussed in the Section 3.1, discrepancies are expected due to the record distributions of the three data sources. The first three most productive DKE authors are *Tan*, *Bhowmick*, and *Manolopoulos*. The ScienceDirect list ranks *van de Riet* and *Storey* and as the 4th and 5th most productive authors, whereas the Scopus list ranks van de Riet as the 13th and Storey as the 6th authors. The WoS list does not include *van de Riet* with more than five DKE papers and ranks *Storey* at the 10th position.

The most active institutions of DKE papers are listed in Table 3. Nanyang Technology University in Singapore appears 10 times in authors' affiliations. The second place is University of Missouri in the USA of 9 times. University College of London, England, is at the third place with eight occurrences. Note that because of the way the data is organized, we cannot conclude that authors from Nanyang Technology University have published 10 DKE papers. Based on the available data we can only state that there are 10 authoring instances in the DKE dataset. The 10 instances could represent 10 distinct DKE papers, or a single DKE paper coauthored by 10 authors all from Nanyang Technology University. Therefore, we regard these institutions as the most active ones, instead of the most productive ones.

Table 4 lists top-10 most cited authors by DKE papers during the period of 1994 through March 20, 2008. Abiteboul takes the first place with the most citations of 155 times, followed by Agrawal with 121 citations and van der Raalst with 90 times. Fig. 2 shows an author co-citation network of the same period of time. The network contains 338 authors cited by the DKE dataset and 544 co-citation links. All 338 authors have at least three citations.

The visualization of the network echoes the ranked list. For example, *Abiteboul* has the largest citation circle. On the other hand, the author co-citation map conveys additional information about how these authors have been cited. The node of *Chen* has a strong purple rim, which means it is a pivotal node in the network with the highest betweenness centrality; in other words, it is strategically important in pulling other nodes together. The citation tree-ring of *Agrawal* shows thick layers of yellow–orange[6] rings, indicating that the majority of citations to *Agrawal* were made in recent years. *Batini* is ranked as the

[6] For interpretation of color in Figs. 1–16, the reader is referred to the web version of this article.

Table 3
Most active institutions of DKE papers

#	# Found in affiliations	Institution	Country
1	10	Nanyang Technol. Univ., Sch. Comp. Engn., Singapore 639798, Singapore	Singapore
2	9	Univ. Missouri, Dept. Comp. Sci., Rolla, MO 65409, USA	USA
3	8	Univ. Coll. London, Dept. Comp. Sci., London WC1E 6BT, England	England
4	7	Univ. Manchester, Dept. Comp. Sci., Manchester M13 9PL, Lancs, England	England
5a	6	Natl. Univ. Singapore, Dept. Comp. Sci., Singapore 117543, Singapore	Singapore
5b	6	Univ. Milan, Dipartimento Sci. Informaz, I-20135 Milan, Italy	Italy
6	5	Worcester Polytech. Inst., Dept. Comp. Sci., Worcester, MA 01609, USA	USA
7a	4	Aristotle Univ. Thessaloniki, Dept. Informat., Thessaloniki 54124, Greece	Greece
7b	4	Eindhoven Univ. Technol., Dept. Technol. Management, NL-5600 MB Eindhoven, Netherlands	Netherlands
7c	4	Univ. Arizona, Dept. Comp. Sci., Tucson, AZ 85721, USA	USA
7d	4	Aristotle Univ. Thessaloniki, Dept. Informat., GR-54006 Thessaloniki, Greece	Greece
7e	4	Australian Natl. Univ., Dept. Comp. Sci., Canberra, ACT 0200, Australia	Australia

Table 4
Most cited authors by DKE papers

Rank	Cites	Authors
1	155	Abiteboul, S.
2	121	Agrawal, R.
3	90	Vanderaalst, W.M.P.
4	78	Kim, W.
5	74	Batini, C.
6	71	Ceri, S.
7	70	Elmasri, R.
8	68	Bertino, E.
9	65	Rumbaugh, J.
10	57	Chen, P.P.S.

Source: Web of Science (1994–3/20/2008).

Table 5
The burstness of keywords assigned to DKE papers

Burst Keywords	Burstness	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07
data mining	6.29																							
ontologies	4.82																							
conceptual schema	4.36																							
object-oriented databases	4.18																							
information retrieval	4.06																							
relational database	3.64																							
clustering	3.58																							
entity-relationship model	3.55																							
deductive database	3.54																							
olap	3.32																							
relational databases	3.32																							

Source: ScienceDirect.

5th most cited author in Table 4. The map shows a strong red rim of *Batini*. This visual attribute indicates that citations to *Batini*'s work abruptly increased during this time interval. The map also shows a few other nodes with relative small citation circles but strong citation burst rates of thick red rims. In an interactive mode of CiteSpace, one can explore such author co-citation maps to identify these three types of authors: landmark authors with large-sized citation rings, pivotal authors with strong purple rims, and rising-star authors with strong red rims of citation burst.

4.1.2. Coauthorship

DKE's coauthorship is depicted in two ways, first as a coauthorship network and then as a geospatial overlay on a world map. According to the 932 ScienceDirect records, there are 2385 DKE authors. A coauthorship network of 1638 DKE authors

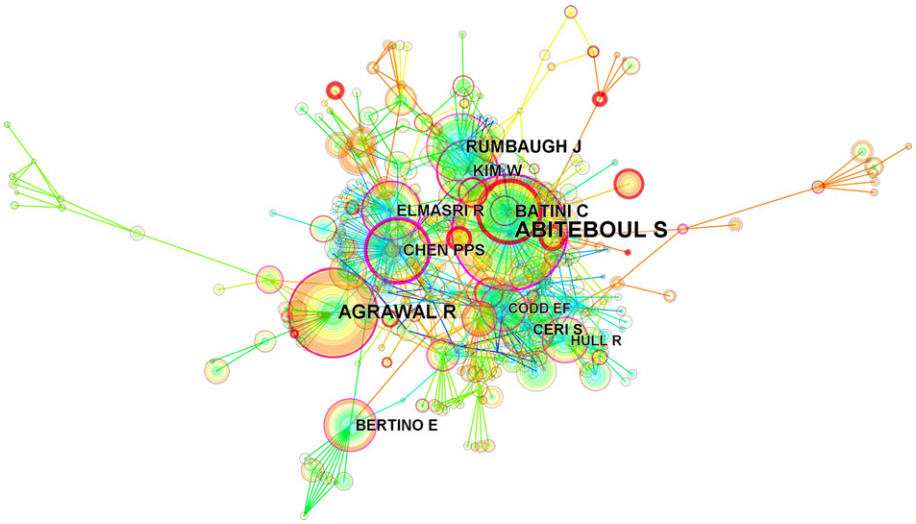


Fig. 2. An author co-citation network (1994–2008), including 338 cited authors and 544 co-citation links. CiteSpace thresholds: 3, 3, 20 throughout.

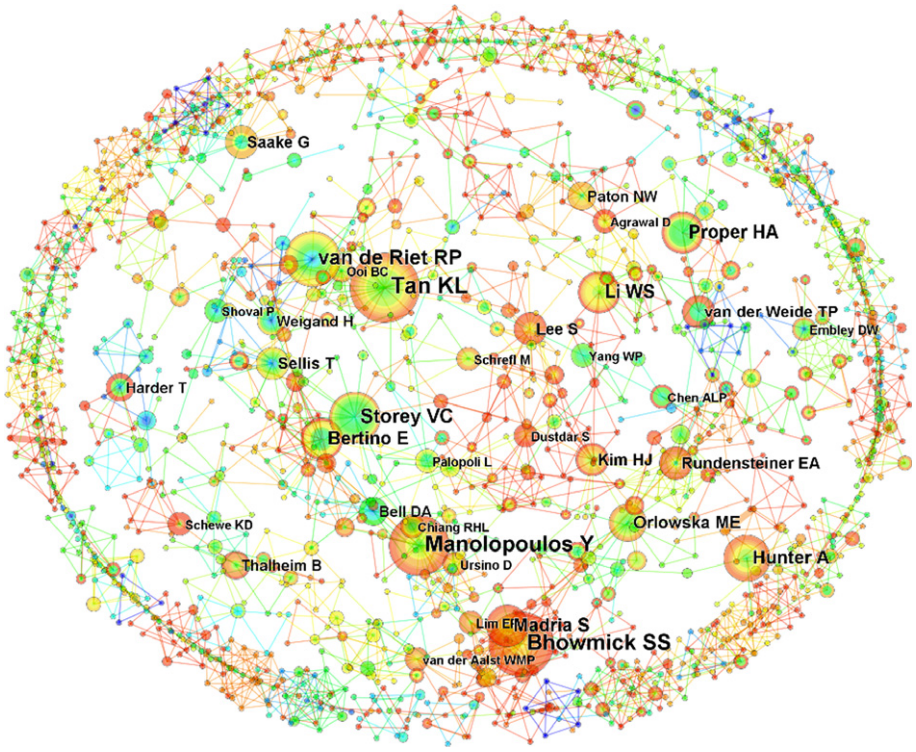


Fig. 3. A coauthorship network of 1638 DKE authors and 2347 coauthoring links based on 932 ScienceDirect records (1985–2007).

of more than one DKE papers and 2347 coauthoring links (see Fig. 3). The legend is the similar to what we have seen so far. The size of a node is proportional to the number of DKE papers one has published. The colors of tree-rings indicate the temporal patterns of a DKE author. For example, the node of *Tan* near the center of the map has the largest citation circle filled by colors from green to yellow and to orange. This pattern indicates that *Tan* has been publishing DKE papers over an extensive period of time. The node for *Storey*, right below the *Tan* node, shows a different pattern. *Storey*'s node is dominated by green citation rings and a thin layer of yellow and orange. This pattern suggests that *Storey* frequently published DKE papers in the green time slices, which correspond to the middle of the 1985–2007 time interval, i.e. about mid 1990s. Then the publication rate declined relatively to the earlier levels of productivity. The colors of lines connecting *Storey* and others echo this

observation – they are essentially in green. In contrast, *Bhowmick*, near the bottom of the map, demonstrates a different pattern. The authorship tree-rings are overwhelmingly yellow and orange, suggesting many of *Bhowmick*'s DKE papers were published in recent 2–3 years.

The coauthorship network is transformed to geospatial overlays on a geographic map of the world (see Fig. 4). The coauthorship network derived from each time slice is mapped to a Google Earth layer. The colors are used to indicate the corresponding time slices. The blue–purple ones indicate earlier time slices, whereas bright red indicates more recent years. The left image in Fig. 4 shows the geographic distribution of DKE authors and their collaborative links in Europe, where there appears to be a high concentration of DKE authors. The middle image in Fig. 4 shows the geographic distribution of DKE authors in the USA. There are slightly more DKE authors on the east coast than elsewhere in the country. The right image shows the distribution in Asia. The star-like pattern in the lower part of the image is centered at the most active DKE institution – *Nanyang Technology University* in Singapore.

Figs. 5 and 6 show the largest and the second largest connected components of the DKE coauthorship network. The number of authors involved in these components underlines the small-world phenomena known as the six-degree association. In Fig. 5, starting from the lower left and moving clockwise, we can see a chain of hubs such as *Proper*, *Orlowska*, *Rundensteiner*, *Bhowmick*, *Lim*, *Storey*, *Kim*, and *Lee*.

Similarly, in Fig. 6, the chain of hubs includes *Jajodia* (upper left), *Castano* (central left) *Bertino* (central), *van de Riet* (upper right) and *Song* (lower right). These hub authors have multiple lines of collaboration with other DKE authors. In this sense, they are the key players and mavens of scientific knowledge in the research community associated to the DKE forum.

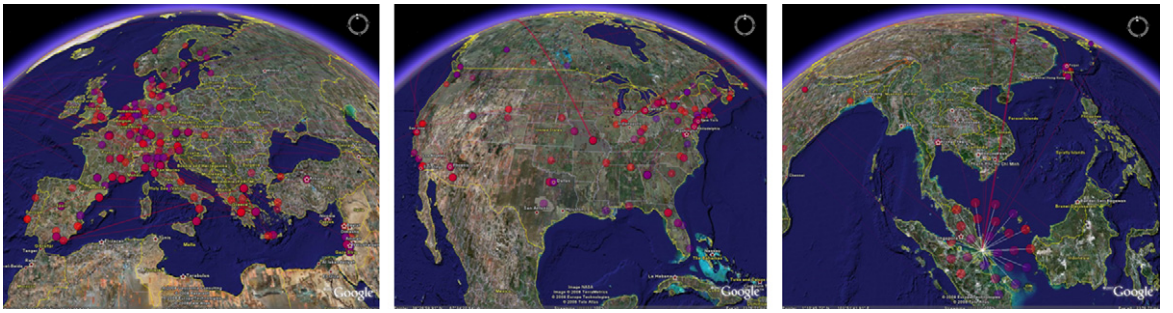


Fig. 4. Geospatial maps with coauthorship network overlays: Europe (left), America (central), and Asia (right). Locations of DKE authors' institutions are marked with blue–red markers, corresponding to the year of publication (blue – earlier; red – recent). Coauthoring links are shown as lines connecting different locations.

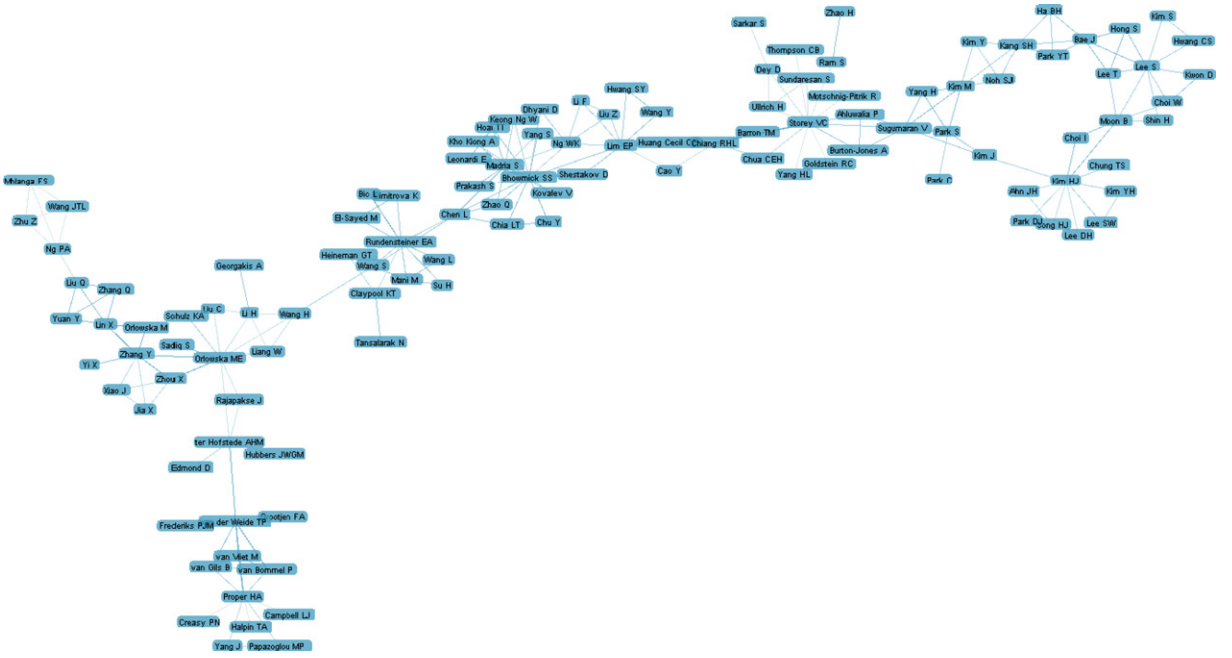


Fig. 5. The largest connected component of the DKE coauthorship network.

4.2. Thematic trends and conceptual structures

4.2.1. Thematic trends

Thematic trends are identified in two ways in terms of burst keywords at macroscopic levels and burst terms at microscopic levels. Since keywords tend to have a higher conceptual level of abstraction than noun phrases, keyword-based burst patterns are expected to identify higher-level trends. The burstness of DKE keywords corresponds to four periods of the entire time span: (1) mid-and late 1980s, (2) early 1990s, (3) early 2000s, and (4) mid- and late 2000s. Burst patterns in the earliest period include *conceptual schema* (1986–1993), *relational database* (1986–1996), and *deductive database* (1986–2002). Burst patterns detected in the second period include *relational databases* (1990–1994), *entity-relationship model* (1991–1996), and *object-oriented databases* (1993–2001). Burst patterns in the third period include *information retrieval* (2000–2004) and *OLAP* (2000–2003). Burst patterns in the most recent period include *data mining* (2005–2007), *ontologies* (2002–2007), and *clustering* (2006–2007).

Fig. 7 plots the timelines of five major burst patterns detected based on the occurrences of DKE keywords. These burst patterns provide useful information about the growth and decay of a specific topic. These timelines provide a more concrete picture of how a topic emerged and faded over time. These burst patterns also provide a useful framework to interpret microscopic-level patterns of burst of noun phrases from free text.

Table 6 shows noun phrases with the most abrupt increases of occurrences in the DKE titles and abstracts. Noun phrases were extracted from the 932 DKE records from the ScienceDirect dataset. The duration of burst is marked by gray blocks for each noun phrase. We will describe these burst patterns in the order of the most recently hot topics, the earliest ones, and the mid-ranged ones.

The sharpest rise goes to the term *xml document*. This is also the most recent burst pattern. Its frequency has been elevated since 2005. Another recent hotspot is *xml data*, which was a particularly popular term for 4 years from 2003 to

Fig. 6. The second largest connected component of the DKE coauthorship network.

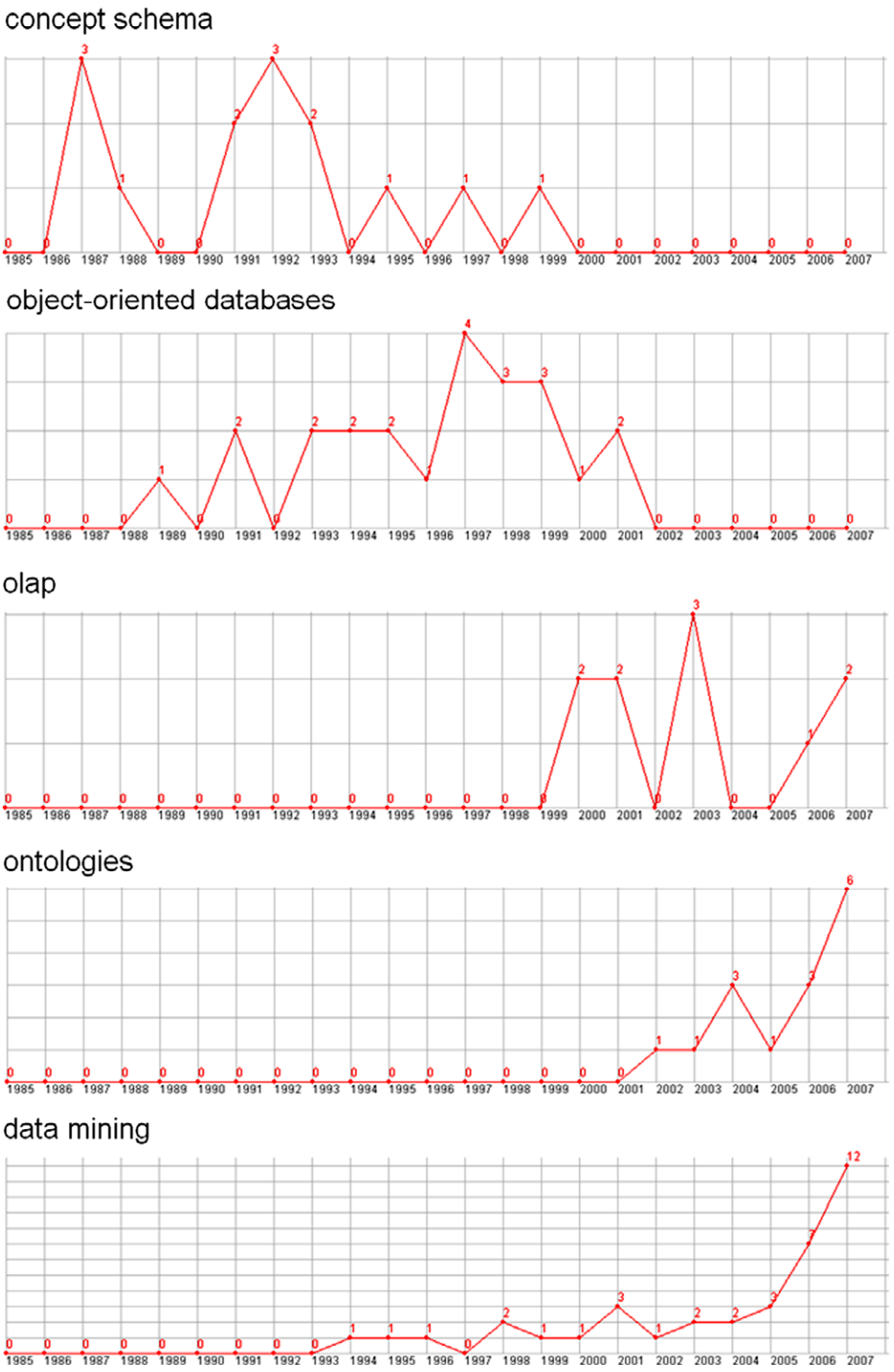


Fig. 7. Macroscopic burst patterns detected over time based on DKE keywords. *Source*: ScienceDirect.

2006. The earliest burst patterns include *conceptual schema* (1986–1997), *deductive database* (1986–1991), and *expert systems* (1985–1993). The mid-ranged ones include *complex objects* (1991–1994) and *conceptual modeling* (1997–2000). Fig. 8 shows a few examples of microscopic-level burst patterns of noun phrases.

Table 6
The burstness of noun phrases extracted from titles and abstracts

Burst Noun Phrases (2-4 Word Phrases)	Burstness	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07
xml document	8.32																							
conceptual schema	5.48																							
complex objects	5.41																							
xml data	4.14																							
database systems	4.00																							
deductive database	3.80																							
expert system	3.73																							
conceptual modeling	3.45																							

Source: ScienceDirect.

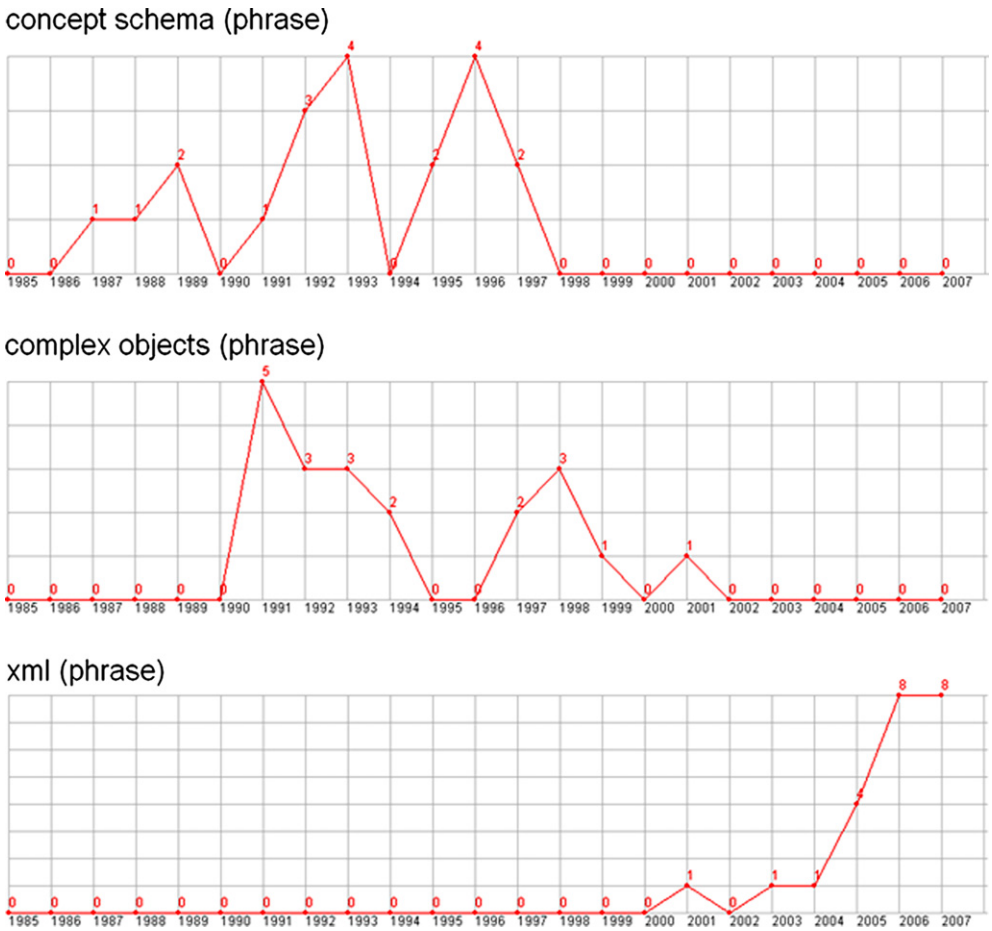


Fig. 8. Microscopic patterns of burst terms.

If *data mining* and *xml document* are fast growing during the same period of time, an interesting question would be whether the co-moving pattern is related. For example, is *xml document* a means to an end of *data mining*? On the other hand, the burst of terms such as *clustering algorithm* would explain the surge of *data mining* papers in DKE.

Fig. 9 shows a hybrid network of keywords and noun phrases. Keywords are shown as circles, whereas noun phrases are shown as triangles. The degree of keyword burst is shown as a red rim of its circle. Similarly, the burst of a noun phrase is

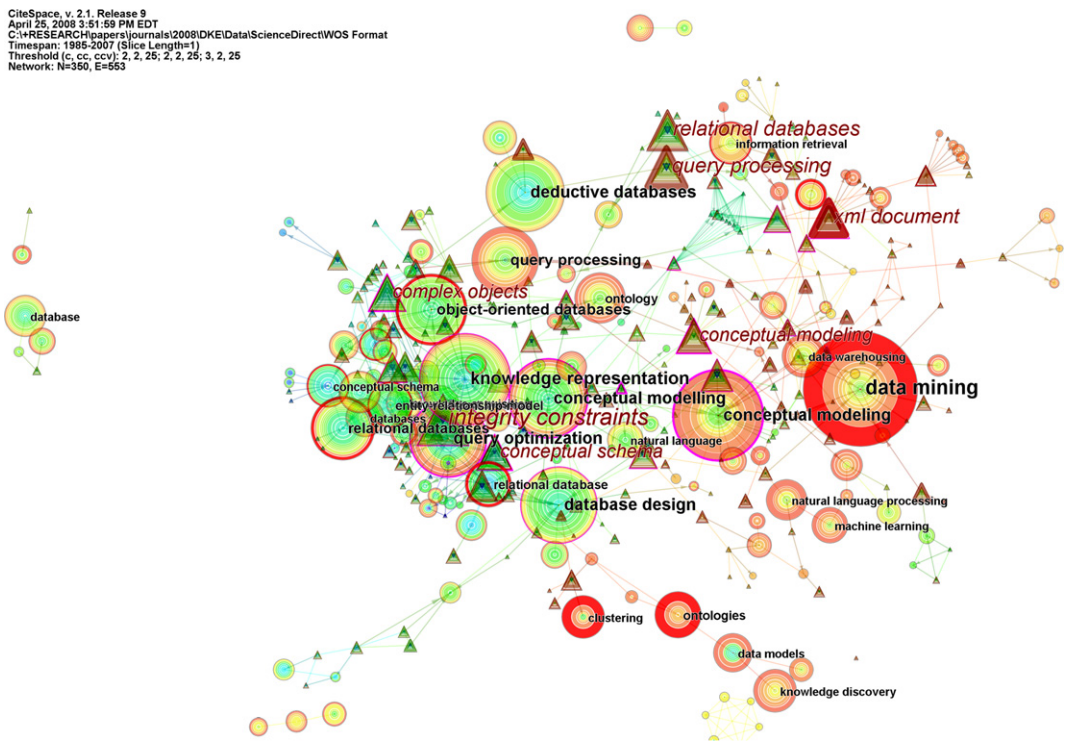


Fig. 9. A hybrid network of keywords (shown as circles with black labels) and noun phrases (shown as triangles with red labels) (1985–2007). The burstness of a node is depicted as a red ring outlined the node. Most keywords in the left-hand size of the map emerged early in the time span, whereas keywords in the right-hand side of the map emerged later in the time span.

shown as a red triangle. The most recent macroscopic burst patterns include ***data mining*** (right), ***ontologies*** (low), and ***clustering*** (low). This network also shows that there is no direct connection between the term *xml document* and the keyword *data mining*. In terms of pivotal nodes, the ones with purple rings, the map shows *conceptual modeling* (middle right), *query optimization* (left), and *knowledge representation* (left).

4.2.2. Conceptual structures

The analysis of the DKE conceptual structures consists of two steps: (1) a clustering analysis based on pLSA on single-word terms, and (2) an analysis of the major latent thematic dimensions based on LSA on multiple-word terms extracted from DKE titles and abstracts over three 8-year periods.

4.2.2.1. Clustering analysis. DKE conceptual structure is based on clusters modeled by pLSA. Table 7 lists 20 clusters identified by pLSA along with the top-5 words selected according to their $p(\boldsymbol{w}|\boldsymbol{z})$ values. The number of clusters is the default setting of the Lemur Toolkit. In fact, we used all default settings of the Lemure toolkit for pLSA clustering. The nature of each cluster becomes clearer when cross-referencing with the most representative papers in these clusters. Table 8 summarizes the 20 clusters with the titles of the two most representative DKE papers. The $p(\boldsymbol{d}|\boldsymbol{z})$ is the probability of seeing a given paper $\boldsymbol{d}$ if the underlying cluster, or the latent variable, is $\boldsymbol{z}$.

Title words are highlighted to suggest the nature of a given cluster with reference to the top-5 words in the same cluster. For example, in Cluster 1, title words **cooperative transactions** and **workflows** are highlighted because they are most similar to the top-5 words by $p(\boldsymbol{w}|\boldsymbol{z})$. While the nature of some of the clusters is straightforward to identify, for example Cluster 5 on high-dimensional data and Cluster 14 on information extraction, the nature of most clusters remains to be ambiguous and diverse. In such situations, the burst detection-based approach appears to have the advantage of identifying specific patterns without the need of prior knowledge of the subject domain.

4.2.2.2. Latent semantic dimensions. The history of DKE is divided into three 8-year periods: (1) 1985–1992, (2) 1993–2000, and (3) 2001–2008. Table 9 lists the frequencies of noun phrases that appeared more than 5 times in respective periods. We removed some generic terms often fount in scientific literature of this type, including *different types*, *different levels*, *new approach*, *new algorithm*, *novel approach*, *recent years*, and *large number*. We also combined singular and plural forms into plural ones, for example *information system(s)*, *deductive database(s)*, *expert system(s)*, *database system(s)*, and *conceptual model(s)*. Different spellings of the same word are also unified, for example, *modeling* and *modelling*. The most

Table 7
Most representative words in 20 clusters generated by probabilistic LSA (pLSA)

#	Size	Term	$p(\boldsymbol{w}\|\boldsymbol{z})$	Term	$p(\boldsymbol{w}\|\boldsymbol{z})$	Term	$p(\boldsymbol{w}\|\boldsymbol{z})$	Term	$p(\boldsymbol{w}\|\boldsymbol{z})$	Term	$p(\boldsymbol{w}\|\boldsymbol{z})$
1	42	Workflow	0.0328	Service	0.0249	Pattern	0.0214	Mine	0.0211	Policy	0.0168
2	38	Rule	0.0150	Model	0.0140	Framework	0.0095	Knowledge	0.0095	Process	0.0088
3	35	Product	0.0310	Table	0.0255	Authorize	0.0192	Knowledge	0.0140	Histogram	0.0116
4	32	Transact	0.0528	Relationship	0.0512	Cardinal	0.0331	Semantic	0.0207	Binary	0.0196
5	31	Privacy	0.0533	Linear	0.0180	Dimension	0.0170	Neighbor	0.0140	Biology	0.0121
6	19	Parallel	0.0438	Couple	0.0262	Stream	0.0205	Time	0.0193	Mpeg	0.0180
7	29	Standard	0.0347	Cover	0.0316	Property	0.0225	Quality	0.0125	Lexicon	0.0107
8	43	Query	0.0800	Answer	0.0348	Graph	0.0177	Set	0.0096	Process	0.0087
9	37	Process	0.0403	Image	0.0343	Retrieval	0.0228	Geometric	0.0118	Model	0.0104
10	27	Secure	0.0694	Inference	0.0372	Multilevel	0.0358	MLS	0.0177	Theorem	0.0116
11	86	Model	0.0627	Conceptual	0.0251	Language	0.0250	Schema	0.0250	Object	0.0130
12	32	Cube	0.0367	OLAP	0.0343	Cell	0.0341	Grid	0.0232	Lattice	0.0161
13	51	Constraint	0.0954	Integrity	0.0457	Rule	0.0347	Event	0.0282	Dependency	0.0236
14	57	Ontology	0.0367	Text	0.0316	Test	0.0121	Knowledge	0.0110	Library	0.0110
15	63	Cluster	0.0186	Algorithm	0.0180	Method	0.0168	Time	0.0143	Performance	0.0142
16	41	Model	0.0243	Knowledge	0.0142	Concept	0.0124	Language	0.0114	Object	0.0106
17	48	Query	0.0264	Xml	0.0190	Algorithm	0.0142	Model	0.0115	Process	0.0083
18	23	Broadcast	0.0610	Client	0.0477	Tune	0.0242	Channel	0.0239	Wireless	0.0181
19	34	Web	0.1287	Page	0.0587	Document	0.0423	HTML	0.0222	Template	0.0177
20	52	Object	0.0165	Design	0.0117	Process	0.0116	Knowledge	0.0105	Schema	0.0094

Domain-specific stopwords such as ***data***, ***database***, **and** ***system***, and general stopwords such as ***paper***, ***base***, **and** ***information are omitted.***

Table 8
Most representative DKE papers in 20 clusters generated by pLSA

Cluster #	$p(d\|z)$ (%)	Paper title
1	3.56	Performance analysis of long-lived **cooperative transactions** in active DBMS By Kangsabanik, P., Yadav, D.S., Mall, R., Majumdar, A.K. Published in 2007
1	2.86	Facilitating cross-organisational **workflows** with a workflow view approach By Schulz, K.A., Orlowska, M.E. Published in 2004
2	2.32	**EDM**: A general framework for Data Mining based on Evidence Theory By Anand, S.S., Bell, D.A., Hughes, J.G. Published in 1996.
2	2.28	Generalized union and project operations for pooling uncertain and imprecise information By Bell, D.A., Guan, J.W., Lee, S.K. Published in 1996
3	8.19	A note on web intelligence, world **knowledge** and fuzzy logic By Zadeh, L.A. Published in 2004
3	4.23	Restructuring decision tables for elucidation of **knowledge** By Hewett, R., Leuchner, J. Published in 2003.
4	10.71	Analysis of binary/ternary cardinality combinations in **entity-relationship modeling** By Jones, T.H., Song, I.Y. Published in 1996
4	6.28	Ownership as a **conceptual modeling** construct By Halper, M., Liu, L.M., Geller, J., Perl, Y. Published in 2007
5	10.17	Array-index: a plug and search K nearest neighbors method for **high-dimensional data** By Aghbari, Z.A. Published in 2005
5	7.17	The Active Vertice method: a performant filtering approach to **high-dimensional indexing** By Balko, S., Schmitt, I., Saake, G. Published in 2004
6	12.25	Warping the time on **data streams** By Capitani, P., Ciaccia, P. Published in 2007
6	10.62	Mapping, indexing and querying of **MPEG-7** descriptors in RDBMS with IXMDB By Chu, Y., Chia, L.T., Bhowmick, S.S. Published in 2007
7	14.31	Candidate interoperability standards: An **ontological overlap analysis** By Green, P., Rosemann, M., Indulska, M., Manning, C. Published in 2007
7	10.24	On cyclic covers and perfect models By Johnson, C.A. Published in 1999
8	3.43	On **querying** simple conceptual graphs with negation By Mugnier, M.L., Leclere, M. Published in 2007
8	3.30	**Query evaluation** in recursive databases: bottom–up and top–down reconciled By Bry, F. Published in 1990
9	4.72	An integrated and collaborative framework for **business design**: a knowledge engineering approach By Seshasai, S., Gupta, A., Kumar, A. Published in 2005
9	4.32	Merging **news reports** that describe events By Hunter, A., Summerton, R. Published in 2006

Table 8 (*continued*)

Cluster #	$p(d\|z)$ (%)	Paper title
10	8.86	Specifying dynamic and deontic **integrity constraints** By Wieringa, R., Meyer, J.J., Weigand, H. Published in 1989
10	7.74	Combining data-driven systems for improving **Named Entity Recognition** By Kozareva, Z., Ferrandez, O., Montoyo, A., Munoz, R., Suarez, A., Gomez, J. Published in 2007
11	1.29	Graph rewriting systems for the **entity-relationship** approach By Breiteneder, C.J., Muck, T.A. Published in 1995
11	1.27	A generic model for 3-dimensional **conceptual modelling** By Creasy, P.N., Proper, H.A. Published in 1996
12	12.13	Cell trees: an adaptive synopsis structure for clustering **multidimensional** on-line **data streams** By Park, N.H., Lee, W.S. Published in 2007
12	11.10	**Load balancing** and data placement for multi-tiered database systems By Li, W.S., Zilio, D.C., Batra, V.S., Zuzarte, C., Narang, I. Published in 2007
13	3.37	Processing production rules in DEVICE, an active **knowledge base system** By Bassiliades, N., Vlahavas, I. Published in 1997
13	3.30	A general treatment of dynamic **integrity constraints** By de Brock, E.O. Published in 2000
14	3.36	Conceptual model-based **data extraction** from multiple-record Web pages By Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng, Y.K., Smith, R.D. Published in 1999
14	3.33	Linguistically based **conceptual modeling** of business communication By Steuten, A.A.G., van de Riet, R.P., Dietz, J.L.G. Published in 2000
15	1.87	Freshness-driven **adaptive caching** for dynamic content Web sites By Li, W.S., Po, O., Hsiung, W.P., Selcuk Candan, K., Agrawal, D. Published in 2003
15	1.21	Algorithms for processing K-closest-pair **queries** in spatial databases By Corral, A., Manolopoulos, Y., Theodoridis, Y., Vassilakopoulos, M. Published in 2004
16	1.30	Analysis of **part-whole relation** and subsumption in the medical domain By Bernauer, J. Published in 1996
16	0.88	An intensional semantics for a hybrid language By Cappelli, A., Mazzeranghi, D. Published in 1994
17	1.18	Faster joins, self-joins and multi-way joins using **join indices** By Lei, H., Ross, K.A. Published in 1998
17	1.18	Faster joins, self-joins and multi-way joins using **join indices** By Lei, H., Ross, K.A. Published in 1999
18	10.42	On selective tuning in unreliable **wireless** channels By Tan, K.L., Chin Ooi, B. Published in 1998
18	9.51	A **dynamic scheduler** for the infinite air-cache By Tan, K.L., Yu, J.X. Published in 1997
19	5.79	Using HMM to learn user **browsing patterns** for focused Web crawling By Liu, H., Janssen, J., Milios, E. Published in 2006
19	4.60	Generation of natural language from information in a frame structure By Perkins, W.A. Published in 1989
20	0.88	Relixpert – an **expert system** shell written in a database programming language By Merrett, T.H. Published in 1991
20	0.72	Weaving temporal and reliability aspects into a schema tapestry By Dyreson, C., Snodgrass, R.T., Currim, F., Currim, S., Joshi, S. Published in 2007

frequently appeared term in the first period is *deductive databases*. The most popular noun phrases in the second period is *relational databases*. The most popular terms in the third period is *xml documents*. These frequency distributions echo the patterns of burst terms identified in earlier sections. According to Table 5, the burst of ***data mining*** was not detected until 2005. In contrast, the presence of *data mining*-related noun phrases was evident in the second period between 1993 and 2000, much earlier than 2005.

Tables 10–12 summarize the results of latent semantic analysis (LSA) of the three periods of data. This analysis is based on noun phrases that appeared at least twice in respective periods. The first dimension is usually the most predominant one but often very diverse as well. Thus, the tables show the top-10 terms along the first dimension. Terms in other dimensions are shown if their projections on the latent variable are greater than or equal to the value of 0.50. The first period is characterized by topics such as *conceptual schema*, *deductive database*, *complex objects*, *enterprise models*, and *natural language constructs*.

The major topics in the second period include *integrity constraints*, *data mining*, *application EER diagram*, *inclusion and concrete dependencies*, and *deductive databases*. Note that the strength of *deductive databases* reduced from the strongest dimension to the fifth and sixth strongest ones. The *data mining* topic, including terms such as *discovery process*, *evidence theory*, and *knowledge discovery* clearly indicates the nature of the second strongest dimension. It is particularly interesting if we compare the burst pattern of the keyword ***data mining*** and the latent topic variable of *data mining* identified by the

Table 9
Frequencies of noun phrases in three 8-year periods (total frequencies > 5)

1985–1992		1993–2000		2001–2008	
21	Deductive databases	19	Relational databases	23	Xml documents
16	Expert systems	18	Conceptual modeling	15	Conceptual modeling
14	Database systems	18	Integrity constraints	14	Xml data
11	Complex objects	14	Information systems	12	Query processing
9	Conceptual schema	15	Object-oriented databases	11	Data mining
9	Integrity constraints	15	Complex objects	10	Conceptual model
8	Relational algebra	13	Conceptual schema	10	Data warehouses
7	Knowledge representation	12	Database design	9	Xml document
6	Knowledge base	12	Database systems	9	Data sources
6	Database design	10	Deductive databases	9	On-line analytical processing
6	Data model	10	Query processing	9	Information extraction
6	Recursive queries	9	Data model	8	Association rules
		9	Relational model	8	Web data
		8	Conceptual models	8	Relational databases
		8	Natural language	7	Knowledge discovery
		7	Knowledge base	7	Information retrieval
		7	Artificial intelligence	6	Web services
		6	Linguistic knowledge	6	Web pages
		6	Schema integration	6	Data model
		6	Knowledge representation	6	Information systems
		6	Knowledge-based systems	6	Object-oriented databases
		6	Data mining		
		6	Data models		
		6	Expressive power		
		6	Object-oriented data models		
		6	Object-oriented database systems		
		6	Query optimization		

method of LSA. Furthermore, according to Table 9, the noun phrase of data mining appeared 6 times in the second period. Therefore, the identification of data mining as one of the major thematic dimensions in the second period is a promising finding.

The concept structure of the DKE papers in the third period consists of latent thematic dimensions such as data mining (*association rules*), cognitive mapping techniques, data access (*cache reusability*, *data access time*, and *communication bandwidth*), multidimensional conceptual modeling (*conceptual multidimensional model*, *multidimensional normal forms*), data warehouses and multidimensional (MD) modeling (*data warehouses*, *main MD properties*, *MD modeling*).

Fig. 10 depicts a minimum spanning tree derived from a hybrid network of keywords and noun phrases. Since keywords tend to represent macroscopic topics and noun phrases represent microscopic ones, the hybrid map is expected to reveal concrete connections between concepts at different levels of granularity. DKE author assigned keywords in the map are labeled with a darker background, whereas noun phrases are labeled with a lighter background. A high-resolution version is

Table 10
10-major latent dimensions derived from the ScienceDirect dataset (1985–1992)

Terms (frequency $\geqslant$ 2 per paper, 1985–1992)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Singular values	*27.44*	*5.85*	*5.15*	*5.00*	*4.74*	*4.36*	*4.18*	*4.12*	*4.06*	*4.00*
Conceptual schema	10.41	3.81								
Deductive database	7.16									
Integrity constraints	7.16									
Database schema	6.25	2.59					1.58			
Object schemas	5.10		2.51							
Expert system shell	5.10		2.51	3.54						
Complex objects					2.49	3.36				
Enterprise model								1.41		
Organizational activities								1.41		
Data abstractions								2.12	1.36	
Global database									1.36	
Database languages										1.41
Logical form										1.41
Natural language										1.41
Natural language constructs										1.41

The dimensionality is reduced from 81 to 62.

Table 11
10-major latent dimensions derived from the ScienceDirect dataset (1993–2000)

Terms (frequency ⩾ 2 per paper, 1993–2000)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Singular values	*41.56*	*9.64*	*7.41*	*6.75*	*6.39*	*6.15*	*5.07*	*5.00*	*4.92*	*4.68*
Data objects	9.23									
Integrity constraints	8.06									
Data mining	8.37	5.71								
General framework		4.63								
Discovery process		3.70								
Evidence theory		3.70								
Knowledge discovery		1.85								
Application data			0.93	4.64						
Default logic	7.08		0.56		1.58	2.13				
Application eer diagram			0.56	2.79						
Cyclic covers			0.52			2.11				
Inclusion dependencies				2.03						
Concrete dependencies				1.86						
Ground facts				1.86						
Active rules	7.06									
Deductive databases					1.58	2.32				
General rules					1.89	2.86				
Conjunctive answers						1.68				
Classical logic						1.43				
Default databases						1.43				
End time							2.03	3.54		
Data schemas							1.63			
Data schema							1.22			
General theory									0.83	
Entity types									0.69	
Application domains									0.62	
Data cube									0.61	2.71
Data cubes									0.61	2.71

The dimensionality is reduced from 209 to 144.

Table 12
10-major latent dimensions derived from the ScienceDirect dataset (2001–2008)

Terms (frequency ⩾ 2 per paper, 2001-2008)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Singular values	*40.22*	*8.83*	*5.96*	*5.48*	*5.41*	*5.30*	*5.00*	*5.00*	*4.88*	*4.72*
Conceptual modeling	13.58	7.09								
Data sources	7.08		0.57			2.26				
Ad hoc processes	6.09									
Association rules	6.00									
Human cognition		3.35								
Cognitive mapping techniques		2.51								
Cache reusability				2.65		0.89				
Data access time				2.65		0.89				
Communication bandwidth				1.77		0.59				
Data items				1.77		0.59				
Long disconnection				1.77		0.59				
Implicational formula					3.02	2.46				
Integrity constraint					2.27	1.85				
Conceptual multidimensional model						1.23				
Multidimensional normal forms						0.82				
Extensional level						0.79				
Data source						0.54				
Extensional integration						0.52				
Dense regions							3.54		1.67	
Conceptual level	5.04							2.22	2.17	
Data warehouses								1.40	1.35	
Main MD properties								1.12	1.12	
MD modeling								1.12	1.12	
Conceptual design								0.53		
Data warehouse								0.53		
Non-local semantics									1.67	
Mediated schema										0.96
Concrete tree types										0.64
Map translation tables										0.64
Mapping relation										0.64

The dimensionality is reduced from 223 to 151.

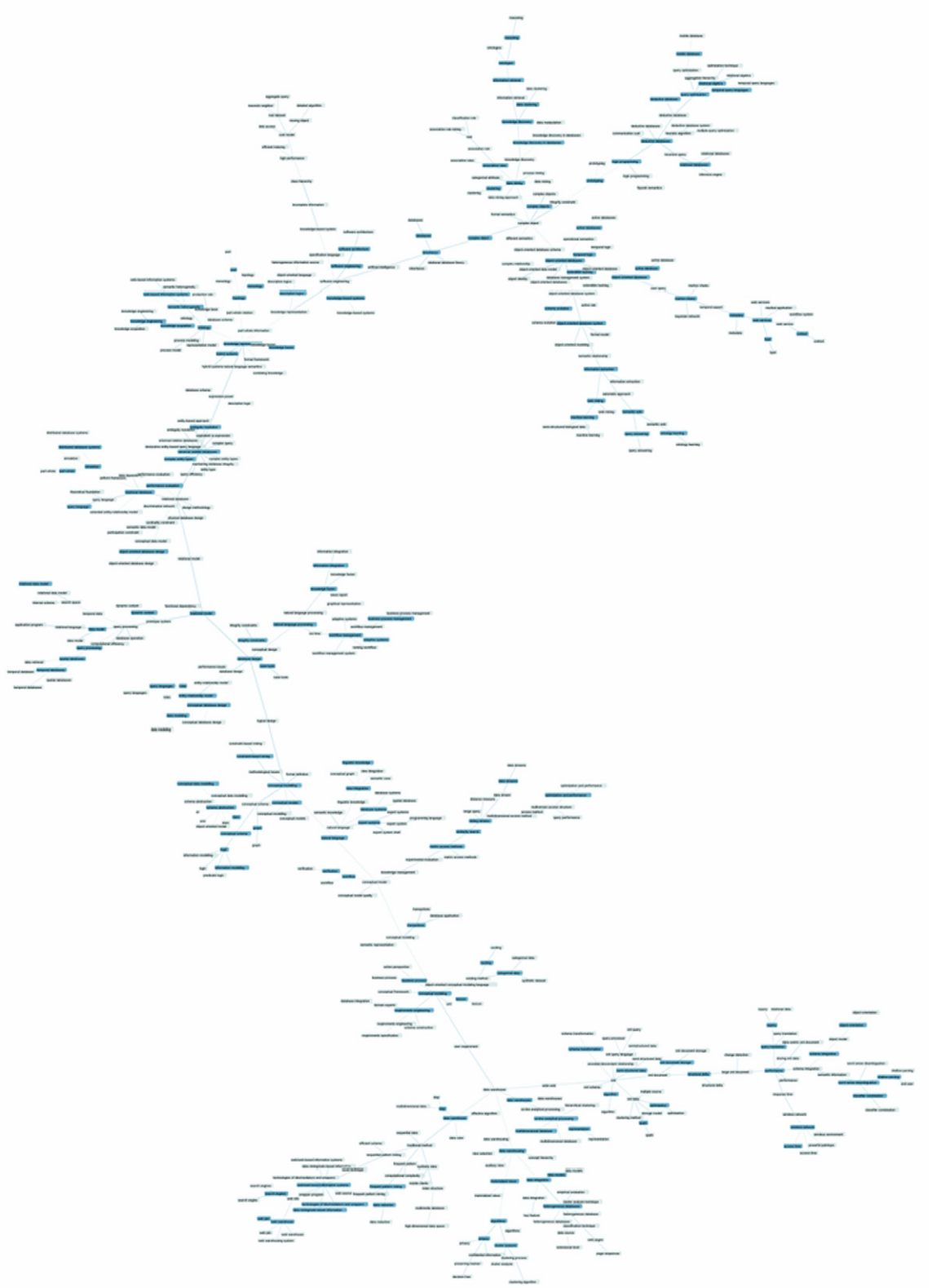


Fig. 10. A concept map of keywords assigned by authors to their own DKE papers (darker labels) and noun phrases extracted from titles and abstracts of DKE papers (lighter labels). This is the greatest component of a minimum spanning tree of 561 nodes and 548 links. CiteSpace thresholds: 2, 2, 20; 2, 2, 20; 3, 2, 20. Source Data: ScienceDirect (1985–2007). The full-size version is available at http://cluster.cis.drexel.edu/~cchen/papers/2008/dke/mst_300dpi_v561e548 _final.png.

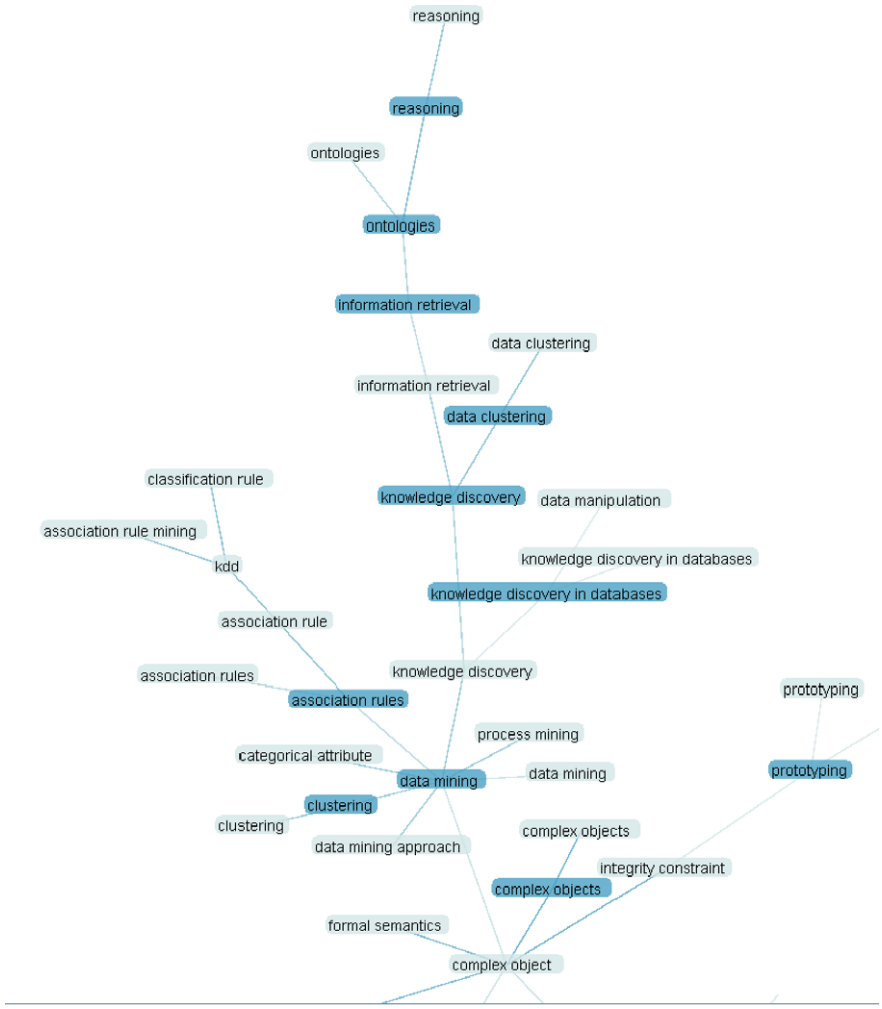


Fig. 11. A *data mining* hub in the minimum spanning tree of the concept map.

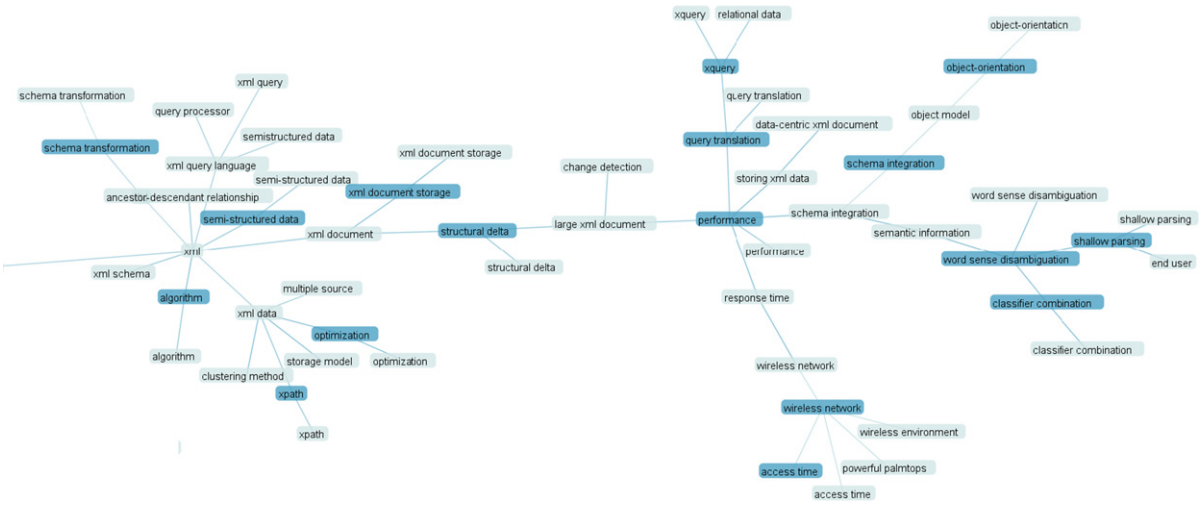


Fig. 12. The *xml* hub in the minimum spanning tree of the concept map.

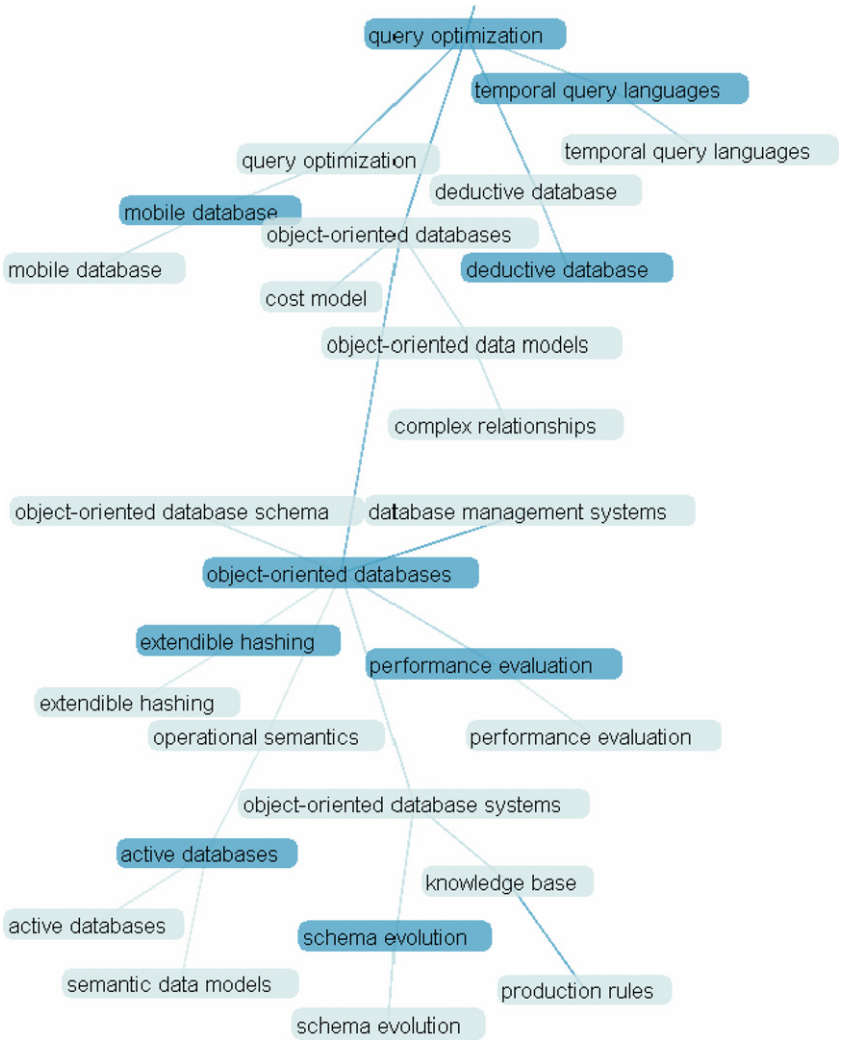


Fig. 13. A sub-network centered on a keyword hub of *object-oriented databases*.

made available on the web. The inclusion of the map is to provide an overall orientation of the conceptual structure of the DKE papers.

Fig. 11 shows a portion of the concept map in Fig. 10. This part of the map includes a hub of keyword ***data mining***. This sub-network is chosen because the burst analysis has identified ***data mining*** as a hot topic in DKE. The hub is connected to other keywords such as ***clustering*** and ***association rules***. The hub is also connected to a number of noun phrases such as *knowledge discovery*, *process mining*, *categorical attributes*, and *data mining approach*. Moving upwards, we come across keywords such as ***knowledge discovery***, ***information retrieval***, and ***ontologies***. This type of concept maps can be useful for learning DKE-specific terminologies.

Fig. 12 shows a sub-network centered on a noun phrase hub *xml*. The sub-graph connects to the rest of the network through the westbound link from *xml* to the term *data warehouse* via the *world-wide web* (not shown in Fig. 12). The hub term is connected to keywords ***algorithm*** and ***semi-structured data*** and to noun phrases such as *ancestor–descendant relationship*, *xml document*, *xml data*, and *xml schema*. Moving across the sub-graph to the right, we found keywords such as ***structural data***, ***performance***, and ***word sense disambiguation*** along the spinal path.

Fig. 13 represents another sub-network with a keyword hub of ***object-oriented databases***. From the earlier sections, we know this is a DKE topic peaked in the mid-range of the time span. The hub is connected to keywords such as ***extendible hashing*** and ***performance evaluation***. It also links to noun phrases such as *object-oriented database schema*, *database management systems*, and *operational semantics*.

Fig. 14 shows a sub-network containing concepts related to entity-relationship models. The keyword ***entity-relationship model*** has a centrality of 48%, which is the 6th greatest one among keywords in the pruned network. If we traverse the sub-network downwards from the hub ***universal relation databases***, we will encounter terms such as *relational database*, which

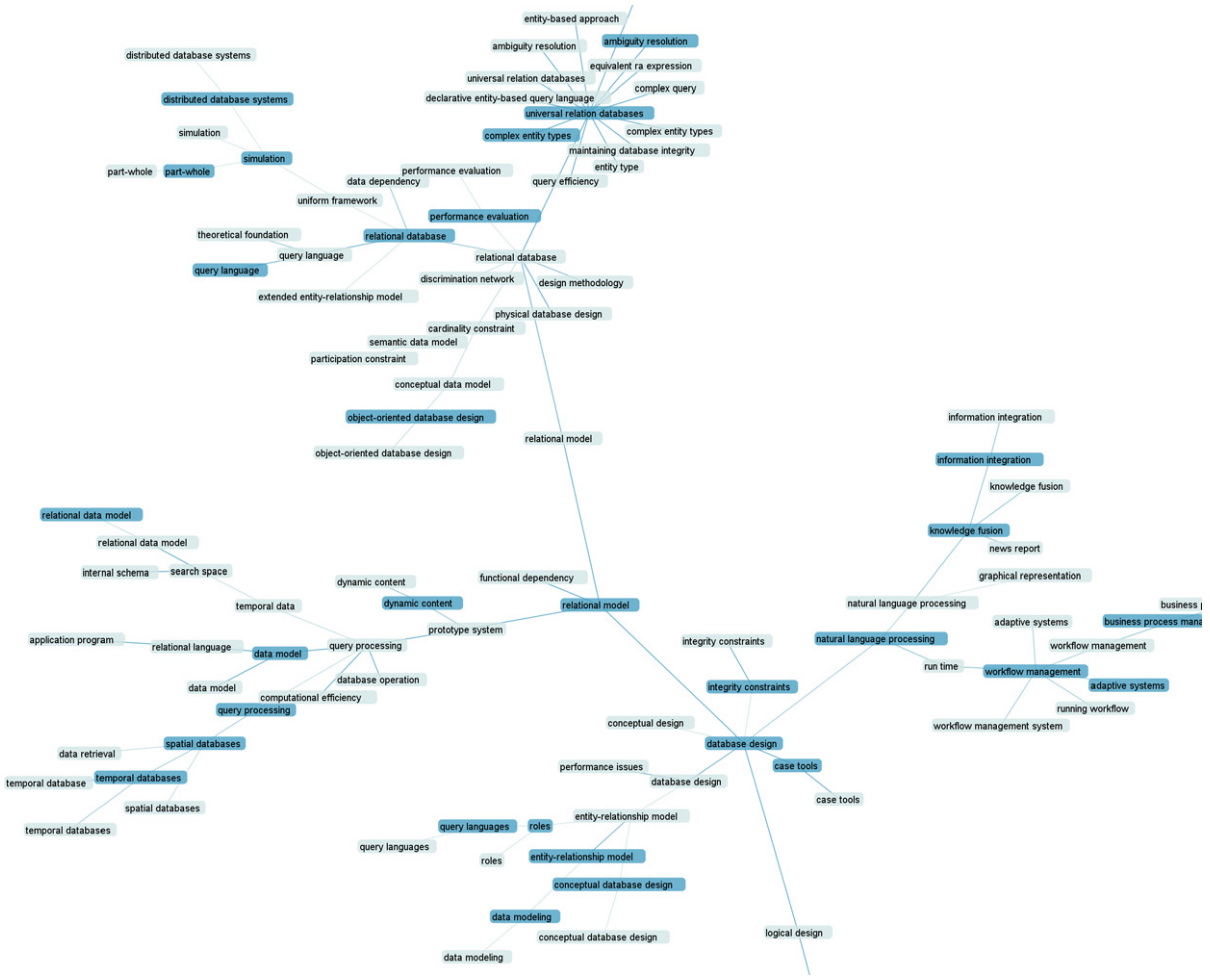


Fig. 14. A sub-network containing concepts related to *entity-relationship models* (lower middle).

links to the term *extended entity-relationship model* in two steps, ***relational model***, and ***database design***. The noun phrase *entity-relationship model* is located in the lower middle branch of the sub-network.

A concept of entity-relationship model has a total of 23 variations as keywords assigned to 42 DKE papers, including *entity-relationship diagram*, *entity-relationship algebra*, *entity-relationship design*, and *entity-relationship data model*. There are 16 variations in noun phrases, including *extended entity-relationship schema*, and *xml entity-relationship exchange*.

4.3. DKE papers

4.3.1. Most cited DKE papers

Table 13 lists the top-20 most cited DKE papers according to two sources of citation data, namely the WoS and Scopus. The citation numbers from Scopus were obtained from the Scopus interface directly. The WoS citations were based on the times cited field (TC) in the WoS dataset. The Scopus counts were based on 3543 records indexed by Scopus, including journal papers and conference papers. The range of the Scopus calculation was 1996–2008, whereas the WoS was based on 1994–2008, more precisely March 20, 2008.

Papers #1–16 are ranked as top-20 by the WoS and Scopus, except #1 and #5 for Scopus. Papers #17–20 are ranked as top-20 most cited in the WoS, but below the top-20s in Scopus. In contrast, papers #21–26 are ranked by Scopus as the top-20s, but ranked lower in the WoS. The main difference between Scopus and the Web of Science is that the Web of Science primarily indexes journal papers, whereas Scopus includes conference proceedings as well as journal papers. A tentative hypothesis would be that papers #17–20 are favorites of journal papers, whereas papers #21–26 are popular in conference papers.

4.3.2. DKE authors' most favorite papers

Table 14 lists papers most frequently cited by DKE authors along with their citation burst rates, betweenness centrality scores in a document co-citation network to be explained shortly, and the citation half life, which is the number of years

Table 13
Top-20 DKE articles most cited in the *Web of Science (WoS)* and *Scopus* (as of March 20, 2008)

#	WoS	Scopus	Title	Author(s)	Year	DKE
1	62	–	Workflow mining: a survey of issues and approaches	van der Aalst, W.M.P.	2003	V. 47, pp. 237
2	58	68	Part-whole relations in object-centered systems: an overview	Artale, A., Franconi, E., Guarino, N., Pazzi, L.	1996	V. 20 (3), pp. 347–383
3	57	81	Workflow evolution	Casati, F., Ceri, S., Pernici, B., Pozzi, G.	1998	V. 24 (3), pp. 211–238
4	55	86	Semantic integration of heterogeneous information sources	Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D.	2001	V. 36 (3), pp. 215–249
5	47	–	SPY-TEC: an efficient indexing method for similarity search in high-dimensional data spaces	Lee, D.H.	2000	V. 34, pp. 77
6	46	45	Mereotopology: A theory of parts and boundaries	Smith, B.	1996	V. 20 (3), pp. 287–303
7	45	59	Supporting ontological analysis of taxonomic relationships	Welty, C., Guarino, N.	2001	V. 39 (1), pp. 51–74
8	42	68	SEMINT: a tool for identifying attribute correspondences in heterogeneous databases using neural networks	Li, W.-S., Clifton, C.	2000	V. 33 (1), pp. 49–84
9	41	57	Conceptual model-based data extraction from multiple-record Web pages	Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng, Y.-K., Smith, R.D.	1999	V. 31 (3), pp. 227–251
10	41	36	Algorithms for inferring functional dependencies from relations	Mannila, Heikki, Raiha, Kari-Jouko	1994	V. 12 (1), pp. 83–99
11	38	39	Parts, wholes, and part-whole relations: the prospects of mereotopology	Varzi, A.C.	1996	V. 20 (3), pp. 259–286
12	37	63	Snoop: an expressive event specification language for active databases	Chakravarthy, S., Mishra, D.	1994	V. 14 (1), pp. 1–26
13	33	61	Computing iceberg concept lattices with TITANIC	Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.	2002	V. 42 (2), pp. 189–222
14	33	46	Reverse engineering of relational databases: extraction of an EER model from a relational database	Chiang, Roger H.L., Barron, Terence M., Storey, Veda C.	1994	V. 12 (2), pp. 107–142
15	29	69	On the representation of roles in object-oriented and conceptual modelling	Steimann, F.	2000	V. 35 (1), pp. 83–106
16	29	48	Building intelligent Web applications using lightweight wrappers	Sahuguet, A., Azavant, F.	2001	V. 36 (3), pp. 283–316
17	29	–	Correctness criteria for dynamic changes in workflow systems: a survey	Rinderle, S.	2004	V. 50, pp. 9
18	26	–	Verification problems in conceptual workflow specifications	ter Hofstede, A.H.M.	1998	V. 24, pp. 239
19	25	–	Argumentative logics: reasoning with classically inconsistent information	Elvanggoransson, M.	1995	V. 16, pp. 125
20	24	–	An overview of the ONIONS project: applying ontologies to the integration of medical terminologies	Gangemi, A.	1999	V. 31, pp. 183
21	–	192	Knowledge Engineering: principles and methods	Studer, R., Benjamins, V.R., Fensel, D.	1998	V. 25 (1–2), pp. 161–197
22	–	47	Information agent technology for the Internet: a survey	Klusch, M.	2001	V. 36 (3), pp. 337–372
23	–	45	Category concept: an extension to the entity-relationship model	Elmasri, R., Weeldreyer, J., Hevner, A.	1985	V. 1 (1), pp. 75–116
24	–	37	The semantic web: yet another hip?	Ding, Y., Fensel, D., Klein, M., Omelayenko, B.	2002	V. 41 (2–3), pp. 205–227
25	–	37	Expressiveness in conceptual data modelling	ter Hofstede, A.H.M., van der Weide, Th.P.	1993	V. 10 (1), pp. 65–100
26	–	36	How to structure and access XML documents with ontologies	Erdmann, M., Studer, R.	2001	V. 36 (3), pp. 317–335

taken for a particular paper to receive half of its current citations. The most cited paper is the 1976 ER paper by *Chen, P.P.S.*, cited by 50 DKE papers. It also has the highest centrality score of 0.78, which underlines the profound impact of the paper on the research community. The one with the most abrupt burst citation pattern is the 1991 book by *Rumbaugh* et al. on object-oriented modeling and design.

A closer examination of citations to Chen's 1976 paper revealed that the citation count of 50 considerably underestimated the real citations due to the diverse range of discrepancies in ways the same reference appeared in DKE papers. We verified the entries listed in Table 15 and found that they were all wrongly entered one way or another and they are all supposed to be the same reference. For example, the initials were entered as P, PP, or PPS. The journal title of *ACM Transactions on Database Systems* was coded in at least four different ways. The year of publication 1976 might be missing. The correct page

Table 14
Papers most frequently cited by the DKE dataset

Frequency	Burst	Centrality	Year	Reference	Half life
50	4.08	0.78	1976	P.P.-S. Chen, The entity-relationship model—toward a unified view of data, ACM Transactions on Database Systems 1 (March) (1976) 9–36	18
39	6.64	0.23	1991	J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, W. Lorensen, Object-Oriented Modeling and Design, Prentice-Hall, Upper Saddle River, NJ, USA, 1991	4
22	3.71	0.04	1992	C. Batini, S. Ceri, S. B. Navathe, Conceptual Database Design: An Entity-Relationship Approach, Benjamin-Cummings Publishing Co., Redwood City, CA, USA, 1991	4
20	4.10	0.11	1994	R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, in: 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, 1994, pp. 487–499	6
20		0.08	1984	Antonin Guttman, R-Trees: a dynamic index structure for spatial searching, in: SIGMOD Conference, 1984, pp. 47–57	16
18	3.51	0.16	1995	S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases: The Logical Level, Addison-Wesley Longman Publishing Co., Boston, MA, USA, 1995	4
17		0.31	1990	Amit Sheth, James Larson, Federated database systems for managing distributed, heterogeneous, and autonomous databases, ACM Computing Surveys 22 (3) (1990) 183–236	7
17	2.82	0.12	1984	J.F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley Longman Publishing Co., Boston, 1984	12
17		0.12	1990	Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, Bernhard Seeger, The R*-tree: an efficient and robust access method for points and rectangles, in: Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, New Jersey, United States, 1990, pp. 322–331	10
16	2.71	0.02	1989	G.M. Nijssen, T.A. Halpin, Conceptual Schema and Relational Database Design: A Fact Oriented Approach, Prentice Hall, Upper Saddle River, NJ, USA, 1989	5

Table 15
Variations of the 1976 paper by Chen, P.P.S

#	Author	Year	Source	Volume	Page
50	Chen, P.P.S.	1976	ACM T DATABASE SYST	V1	P9
1	Chen, P.P.	–	ACM TODS	V1	P9
1	Chen, P.P.	1976	ACM TODS	V1	
2	Chen, P.P.	1976	ACM T DATA BASE SYST	V1	
3	Chen, P.P.	1976	ACM T DATABASE SYST	V1	P1
1	Chen, P.P.	1976	ACM T DATABASE SYST	V1	P166
1	Chen, P.P.	1976	ACM T DATABASE SYST	V1	
1	Chen, P.P.	1976	ACM T DATABASE SYST	V1	P471
1	Chen, P.	1976	ACM T DATABASE SYST	V1	P96
2	Chen, P.P.	1976	ACM T DATABASE SYST	V1	
1	Chen, P.P.	1976	ACM T DATABASE SYST	V1	P2
1	Chen, P.	1976	ACM T DATABASE SYST		P9
65					

numbers are pages 9–36, but we see P1, P2, as well as the correct P9. Furthermore, pages P96, P166, and P471 were nowhere found in the transactions' volume 1. A more accurate citation count should be 65. In this study, however, we decided not to take the extra data cleaning step. Thus, the numbers shown in this paper are lower bounds of true numbers.

4.3.3. Thematic trends by citation

Citation bursts of papers provide concrete indicators of emerging themes as well as themes that were once highly active to DKE authors. Table 16 lists the burst rates of references cited by DKE papers in chronological order based on the WoS dataset (1994–2007). The pioneering paper by Chen, P.P.S. in 1976 was found to have a 5-year time span of citation burst between 1994 and 1998. The second earliest item on the list is Salton's information retrieval book published in 1983. It is interesting to note it was not until 2006 a citation burst was detected, suggesting a new and ongoing information retrieval trend in DKE. Citations to Rumbaugh et al.'s book on object-oriented modeling and design were peaked during 1995–1997. The most recent strong citation trend is associated with Agrawal and Srikant's 1994 VLDB paper on fast algorithms for finding association rules. The pattern was detected since 2005, which echoes the data mining theme identified by our analysis of conceptual structures with the burst patterns of keywords and noun phrases as well as latent semantic dimensions identified by LSA and pLSA.

Detailed year-by-year citation timelines of three papers with significant burst patterns are shown in Fig. 15. Thicker sections of the lines highlight the duration of burst. The burst durations of Chen (1976) and Rumbaugh et al. (1991) were both located in the early 1990s, whereas the burst of Agrawal and Srikant (1994) is still rising as of the end of 2007, suggesting that the data mining trend is still going strong.

Table 16
References with the strongest citation burstness based on the Web of Science records (1994–2007)

References	Year	Burst	Begin	End	Span	Half life
P.P.-S. Chen, The entity-relationship model—toward a unified view of data, ACM Transactions on Database Systems 1 (March) (1976) 9–36	*1976*	*4.08*	*1994*	*1998*	*5*	*18.0*
G. Salton, Introduction to Modem Information Retrieval, McGrawHill Book, New York, 1983	1983	2.99	2006	2007	2	23.0
J. F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley Longman Publishing Co., Boston, 1984	1984	2.82	1996	1998	3	12.0
T.J. Teorey, D. Yang, J.P. Fry, A logical design methodology for relational databases using the extended entity-relationship model, ACM Computing Surveys 18 (1986) 197–222	1986	4.25	1994	1999	6	10.0
J.W. Lloyd, Foundations of Logic Programming, Springer-Verlag, New York, NY, 1987	1987	2.82	1995	1997	3	8.0
J. Peckham, F. Maryanski, Semantic data models, ACM Computing Surveys 20 (1988) 153–189	1988	3.09	1995	1999	5	7.0
G.M. Nijssen, T.A. Halpin, Conceptual Schema and Relational Database Design: A Fact Oriented Approach, Prentice Hall, Upper Saddle River, NJ, USA, 1989	1989	2.71	1994	1997	4	5.0
J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, W. Lorensen, Object-Oriented Modeling and Design, Prentice-Hall, Upper Saddle River, NJ, USA, 1991	*1991*	*6.64*	*1995*	*1997*	*3*	*4.0*
P. van Bommel, A.H.M. ter Hofstede, T.P. van der Weide, Semantics and verification of object-role models, Information Systems 16 (1991) 471–495	1991	2.74	1994	1997	4	3.0
C. Batini, S. Ceri, S. B. Navathe, Conceptual Database Design: An Entity-Relationship Approach, Benjamin-Cummings Publishing Co., Redwood City, CA, USA, 1991	1992	3.71	1996	1998	3	4.0
A.H.M. Ter Hofstede, H.A. Proper, T. P. van der Weide, Formal definition of a conceptual language for the description and manipulation of information models, Information Systems 18 (1993) 489–523	1993	3.05	1994	1996	3	1.0
R. Elmasri, S.B. Navathe, Fundamentals of Database Systems, Benjamin-Cummings Publishing Co., Redwood City, CA, USA, 1994	1994	4.25	1996	1998	3	2.0
R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, 1994, pp. 487–499	*1994*	*4.10*	*2005*	*2007*	*3*	*6.0*
S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases: The Logical Level, Addison-Wesley Longman Publishing Co., Boston, MA, USA, 1995	1995	3.51	1999	2003	5	4.0
S. Abiteboul, D. Quass, J. McHugh, J. Widom, J. L. Wiener, The Lorel query language for semi-structured data, International Journal on Digital Libraries 1 (1997) 68–88	1997	2.72	2001	2003	3	5.0
M. Reichert, P. Dadam, Adeptflex—supporting dynamic changes of workflows without losing control, Journal of Intelligent Information Systems 10 (1998) 93–129	1998	3.31	2004	2005	2	6.0
R. Goldman, J. McHugh, J. Widom, From semi-structured data to XML: migrating the lore data model and query language, in: The 2nd International Workshop on the Web and Databases, Philadelphia, Pennsylvania, USA, 1999	1999	2.90	2003	2007	5	4.0
W. van der Aalst, K. van Hee, Workflow Management: Models, Methods, and Systems, The MIT Press, Boston, MA, USA, 2002	2002	2.87	2003	2005	3	2.0

4.3.4. Document co-citation networks

Fig. 16 shows a document co-citation network derived from the collective citing behavior of the DKE authors as a whole. The network was generated based on the WoS dataset (1994–2007). It consists of 646 papers that have been cited by two or more DKE papers. It also contains 3945 co-citation links. Each co-citation link between references A and B represents at least two co-citation instances of the pair. Citations made in earlier years are shown in blue and green rings, mid-range years in yellow, and recent years in light brown and orange. Similarly, the colors of co-citation links depict the earliest year in which the connection was made for the first time. For example, it is quite possible that papers published in the 1980s were not co-cited until 1990s. Nodes with red rings are papers with strong citation burst patterns. Nodes with purple rings are known as pivotal-point papers because they have high betweenness centrality scores. They are the brokers or bridges that connect different parts of the network together. Once again, notable papers such as the Chen's 1976 paper, Rumbaugh et al.'s 1991 book, and Agrawal's 1994 paper are featured prominently in the visualized network. In addition to the large and highly connected area in the center of the network, there are more than dozen of distinct clusters. The colors of these clusters indicate when they are mostly cited in DKE papers. For example, we can see a dense cluster in orange right next to the red circle of Agrawal's 1994 paper. Since we know that Agrawal's paper is experiencing a period of burst since 2005, thus the orange cluster identifies the group of papers that form the basis of the trend. We can further predict that this is likely to be a cluster of data mining papers. The visualization also provides a visual confirmation of the breadth and depth of the DKE topics over the years.

5. Discussion and conclusion

We have analyzed the structure and dynamics of thematic trends, semantic clusters, and citation networks of DKE papers (1985–2007). We have examined the 24-year history of DKE from multiple perspectives through a wide range of units of analysis and interrelationships across different types of units:

- Citing authors as contributors to DKE.
- Cited authors as people whose work has influenced DKE authors.
- Author assigned DKE paper keywords as content descriptors at macroscopic levels.
- Noun phrases extracted from the titles and abstracts of DKE papers as the potential indicators of contents.

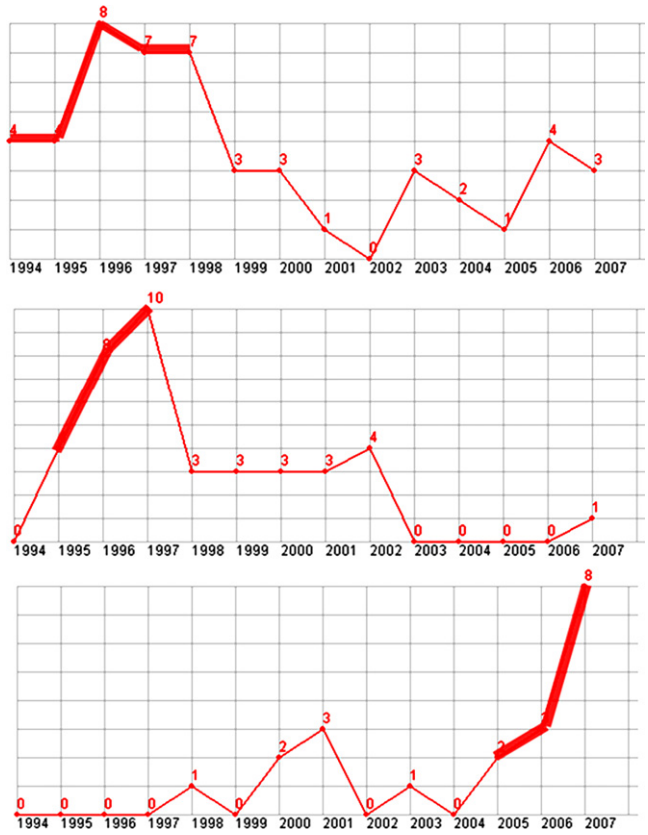


Fig. 15. The citation histories of *Chen (1976)* (top), *Rumbaugh (1991)* (middle), and *Agrawal (1994)* (bottom). Thicker lines indicate the periods of citation burst.

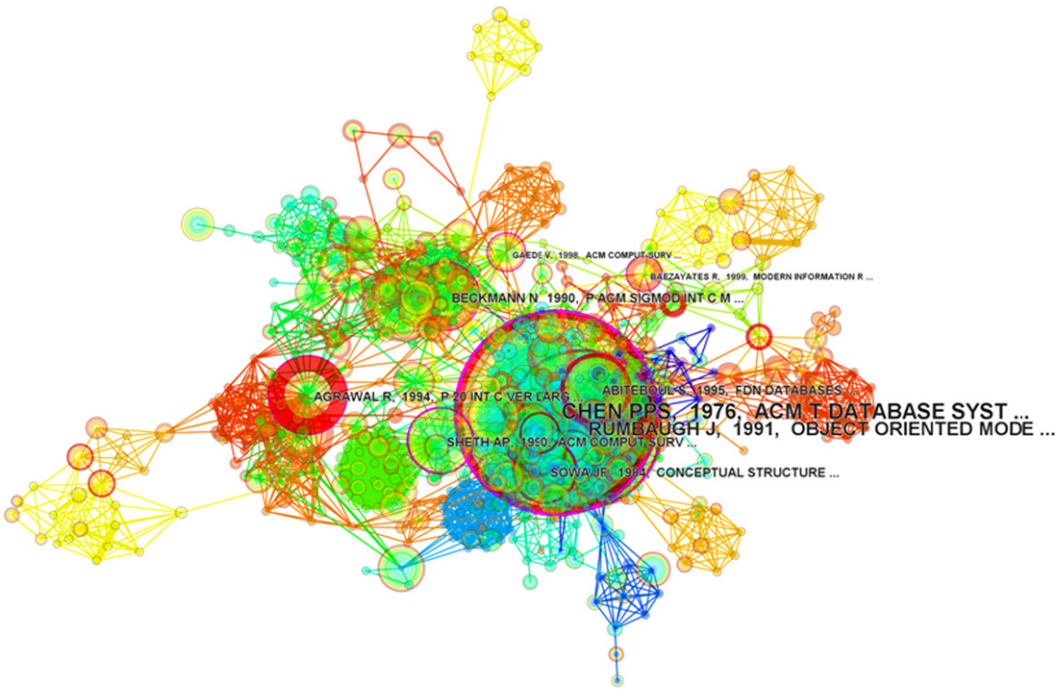


Fig. 16. A document co-citation network of DKE (1994-2007), including 646 papers and 3945 co-citation links. CiteSpace thresholds: 2, 2, 20; 2, 2, 20; 2, 2, 20. References with high centrality scores are labeled. Citation bursts are depicted as red rings.

- Clusters of single- and multiple-word terms.
- Clusters of DKE papers.
- Clusters of author assigned keywords.
- Relationships between noun phrases and author assigned keywords.
- Papers published in DKE.
- Papers cited by DKE papers.

We have identified a variety of thematic trends. Some of them are still going strong, whereas others peaked years ago. We have also identified structural and temporal patterns across multiple levels of aggregation, ranging from terms in text to papers and to clusters of papers. In terms of methodologies, we have found that noun phrase-based analyses detected more timely patterns than author assigned keywords.

We highlight some major conclusions about DKE as follows. The contributing research community to DKE is a well-connected social network containing long chains of collaborating authors. The 1976 entity-relationship model paper by Chen has a central place in the DKE citation image, cited by 65 DKE papers. The concept of entity-relationship modeling is an integral part of the conceptual structure. DKE has a diverse and dynamic landscape of topics and trends. Over the last 24 years, the journal has provided a forum for the forefront of research. The forum is rich in persistent and transient trends. For example, DKE is evidently revealing some emerging trends in areas of data mining and ontologies. There is also an XML-centered trend in connection to topics such as the World Wide Web and data warehousing. In addition, our analysis has identified trends that peaked in the past, including deductive databases, relational databases, and object-oriented databases. DKE has provided its research community a vibrating and stimulating platform for the last 24 years. It gives researchers a unifying forum to showcase their best work in fields concerning data and knowledge engineering.

Acknowledgements

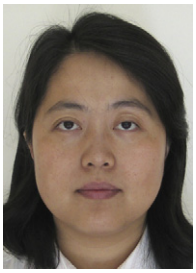
This work is in part supported by the National Science Foundation under Grant No. 0612129. The authors would like to thank for Elsevier and in particular Patrick Gibbons for arranging the access to Scopus and ScienceDirect.

References

[1] C. Chen, CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature, J. Am. Soc. Inform. Sci. Technol. 57 (2006) 359–377.
[2] C. Chen, Searching for intellectual turning points: progressive knowledge domain visualization, Proc. Natl. Acad. Sci. USA 101 (2004) 5303–5310.
[3] J. Kleinberg, Bursty and hierarchical structure in streams, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002, pp. 91–101.
[4] R.W. Schvaneveldt, Pathfinder associative networks: studies in knowledge organization, in: D. Partridge (Ed.), Ablex Series in Computational Sciences, Ablex Publishing Corporations, Norwood, New Jersey, 1990.
[5] C. Chen, Visualising semantic spaces and author co-citation networks in digital libraries, Inform. Process. Manage. 35 (1999) 401–420.
[6] D. Crane, Invisible Colleges: Diffusion of Knowledge in Scientific Communities, University of Chicago Press, Chicago, Illinois, 1972.
[7] M. Newman, The structure of scientific collaboration networks, in: Natl. Acad. Sci., USA, 2001, pp. 404–409.
[8] D. Price, A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inform. Sci. 27 (1976) 292–306.
[9] T. Hofmann, Probabilistic latent semantic analysis, in: Uncertainty in Artificial Intelligence (UAI'99), Stockholm, Sweden, 1999.

Chaomei Chen is associate professor in the College of Information Science and Technology at Drexel University. He received his bachelor degree in mathematics from Nankai University, China, his master's degree in computation from the University of Oxford, and his doctorate in computer science from the University of Liverpool. His research interests include information visualization, knowledge domain visualization and mapping scientific frontiers. He is the author of *Information Visualization: Beyond the Horizon and Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. He is the founder and the Editor-in-Chief of *Information Visualization* journal.

Il-Yeol Song is a professor of the College of Information Science and Technology at Drexel University. He received the M.S. and Ph.D. degrees from the Department of Computer Science, Louisiana State University, in 1984 and 1988, respectively. His research interests include conceptual modeling, object-oriented analysis, data warehousing, and bioinformatics. He has published over 160 papers. He has won three teaching awards from Drexel University: *Exemplary Teaching Award* in 1992, *Teaching Excellence Award* in 2000, and the *Lindback Distinguished Teaching Award* in 2001. He is a co-author of the *ASIS Pratt_Severn Excellence in Writing Award* at National ASIS meeting (1997), the *Best Paper Award* of in the 2004 IEEE CIBCB 2004. He won 14 research awards from competitions of annual Drexel Research Days. He is a Co-Editor-in-Chief of the *Journal of Computing Science and Engineering (JCSE)*. He is also an associate editor for the *Journal of Database Management*, *International Journal of E-Business Research*, and *Journal of Digital Forensics, Security and Law*. In addition, he was also a guest editor for *Journal of Database Management*, *Data and Knowledge Engineering*, and *Decision Support Systems*. He served as a program/general chair of 16 international conferences/workshops including DOLAP98, CIKM99, ER03, DAWAK07 and DAWAK08.



Xiaojun Yuan is an assistant professor in Department of Information Studies, College of Computing and Information, at University at Albany, State University of New York. She holds a Ph.D. degree in information science from Rutgers, The State University of New Jersey. Her research interests include information retrieval, human information behavior, information visualization, human computer interaction, and user interface design and evaluation.

Jian Zhang is currently a doctoral student at the College of Information Science and Technology, Drexel University, Philadelphia, USA. Before going to Philadelphia, he was an Aggie at Texas A& M University, where he received the MS in Science and Technology Journalism. He grew up in China, where he received his BS in Computer Science in 1997. Then he started to work as an editor, later the vice editor in chief, for a leading Chinese science magazine till 2003.