Scientometrics (2015) 103:471–487
DOI 10.1007/s11192-015-1539-0

Combining OLAP and information networks for bibliographic data analysis: a survey

Sabine Loudcher · Wararat Jakawat · Edmundo Pavel Soriano Morales · Cécile Favre

Received: 9 June 2014 / Published online: 17 February 2015

Abstract In the context of scientometrics and bibliometrics, several research fields are dealing with bibliographic data. In this paper, we will explore how the combination of online analytical processing (OLAP) analysis and information networks could be an interesting issue. In Business Intelligence, OLAP is a technology supported by data warehousing systems. It provides tools for analyzing data according to multiple dimensions and multiple hierarchical levels. At the same time, several information networks (co-authors network, citations network, institutions network, etc.) can be built based on bibliographic databases. Originally, OLAP was introduced to analyze structured data. However, in this paper, we wonder if, by combining OLAP and information networks, we can provide a new way of analyzing bibliographic data. OLAP should be able to handle information networks and be also useful for monitoring, browsing and analyzing the content and the structure of bibliographic networks. The goal of this survey paper is to review previous work on OLAP and information networks dealing with bibliographic data. We also propose a comparison between traditional OLAP and OLAP on information networks and discuss the challenges OLAP faces regarding bibliographic networks.

Keywords OLAP · Information networks · Bibliographic data

S. Loudcher (✉) · W. Jakawat · E. P. S. Morales · C. Favre
Université de Lyon, (ERIC LYON 2), Lyon, France
e-mail: sabine.loudcher@univ-lyon2.fr

W. Jakawat
e-mail: wararat.jakawat@univ-lyon2.fr

E. P. S. Morales
e-mail: edmundo.soriano-morales@univ-lyon2.fr

C. Favre
e-mail: cecile.favre@univ-lyon2.fr

Introduction

Scientometrics and bibliometrics have become standard tools of science policy and research management (Van Raan 1997). Multiple research fields are concerned with bibliographic data analysis (Statistics, Data Mining, Graph Theory, OLAP analysis, etc.) in order to achieve different objectives (relationship studying, ranking, community mining, prediction, etc). This kind of analysis relies on information designed and stored in bibliographic databases. Bibliographic databases contain publications from conference proceedings, journals, books, etc., and store a collection of their fundamental information such as title, authors, year, venue, references and citations. Users can access them online thanks to digital libraries.

Among the different research fields interested in bibliographic data analysis, in this paper we focus on OLAP (Online Analytical Processing) analysis. OLAP is part of the Business Intelligence set of techniques. It can help managers, universities, governments, etc., to take decisions more easily, such as which projects or researchers should receive more support, who should be a reviewer in a journal or a conference, etc. OLAP is a multidimensional data analysis system that provides fast analysis for decision making within a vast amount of data (Chaudhuri and Dayal 1997). Data is organized around indicators (called measures) and analysis axes (called dimensions). Dimension attributes can either form a hierarchy or be purely descriptive. These hierarchies make it possible to obtain views of the data at different granularity levels, i.e., summarized or detailed through the use of *roll-up* and *drill-down* operations respectively. Ferrara and Salini (2012) found interesting the use of a multidimensional approach to bibliographic data analysis. They introduced a multidimensional model for bibliographic data and defined ten challenges that must be addressed in bibliometrics: a conceptual multidimensional model, data availability and integration, duplicate detection and data normalization, data aggregation, comparison and ranking, aggregation of indexes, multiple measures, extraction and indexing of textual data, topic-based analysis of textual data and combining multidimensional information.

Moreover, bibliographic databases can be seen as information networks. In a network there are several types of objects which are interconnected by relationships. From a bibliographic database, we can build a network of authors, a network of citations, a network of conference, etc. As bibliographic networks can include multidimensional attributes, Zhao et al. (2011) spoke about multidimensional information networks, but in general we speak about heterogeneous information networks. Traditionally, OLAP was used to analyze structured data but with the rapid spread of information networks, it must adapt to manage heterogeneous information networks. OLAP on information networks can be useful for monitoring, browsing and analyzing the content and the structure of networks. We want to study what OLAP can bring to the analysis of bibliographic networks and we want to investigate how OLAP should be adapted to deal with information networks, specially with bibliographic networks. In a previous paper, we have proposed a new framework to deal with bibliographic data (Jakawat et al. 2013). The goal of this present survey paper is different. Our goal is to review more works dealing with this topic. We also propose a comparison between traditional OLAP and OLAP on information networks, also called Graph OLAP or Social OLAP. This comparison and the literature review allow us to discuss the challenges OLAP faces while working with bibliographic networks.

The reminder of this paper is organized as follows. In "Bibliographic data analysis" section, we introduce the goals of bibliographic data analysis and the different approaches to tackle them. After recalling basic concepts on information networks and OLAP analysis, we propose in "OLAP on information networks" section a comparison between traditional

OLAP and Graph OLAP. In "Literature review" section, we present the literature review. Discussion section concludes this paper with a discussion and some research issues.

Bibliographic data analysis

Multiple research fields are interested in bibliographic data analysis because it can yield very rich and useful information. This is not an easy task though. Due to the quantity and the variety of approaches that are concerned with this subject, all of them with various goals, it is not possible to provide a comprehensive summary for all of them. Still, in this section, we introduce multiple examples of research goals in bibliographic data analysis and we review relevant existing works in related research fields.

Goals and related fields

In the analysis of bibliographic data, several objectives are interesting :

1. *Search engine* These tools help users to find information about relevant papers (according to authors, conferences, topics, etc.) in order to prepare reports and documentation.
2. *Relationship studying* The structure of bibliographic data is also interesting while studying relationships among entities. Each publication is composed of author(s), venue and related data. Making use of this information, researchers have analyzed the patterns of co-authorship collaborations, the centrality, the structured links between universities and the professional relationship between authors (professor/student relations), among others.
3. *Literature-based discovery* Publications (the literature) can be used to find new relationships between existing knowledge. These findings do not generate new knowledge. Instead, they seek to connect existing knowledge from empirical results by shedding light on neglected relationships.
4. *Ranking* Ranking analysis can be used to assess research. It evaluates objects based on mathematical functions and it compares those of the same type. Several approaches (e.g., impact factors) have been proposed to rank journals, conferences, and authors.
5. *Community mining* The goal is to find groups of objects that share similar properties and that are connected to each other. Identifying these connections and locating objects in different communities is valuable for numerous tasks. For example, to find potential collaborators for researchers, to discover communities in author-conference social networks, to find reviewers to be invited as program committee members, etc.
6. *Topic detection* Its goal is to identify topics by exploring and organizing the content of textual data and to automatically aggregate its information into clusters. In the context of publications, topic detection can cluster publications according to their content, it can find the main discussed topics in a group of conferences, it can also detect the most relevant trends in a given research field, among other tasks.
7. *Multidimensional exploration* Bibliographic databases contain extensive amounts of data. Still, users need only consistent and valuable information, such as portion of objects, links or sub-networks. However, bibliographic data features cannot be taken into account separately. Thus, bibliographic data analysis must be able to support multidimensional exploration and reporting. For instance, it could be useful to follow up, over time, the evolution of discovered topics from a keyword search.

8. *Prediction* Multiple applications for bibliographic network analysis are focusing on predicting links or interactions among objects. A supervised model is used to learn the relations' history. Then, it is able to predict new information such as research trends over time, or in a given community, the emergence of new topics/conferences.

To achieve these goals, various methods can be used; they come from different research fields such as:

- *Statistics* The application of mathematics and statistical methods to bibliographic data analysis is not new. It started in the 1920s and became more popular during the sixties (Hulme 1923; Pritchard 1969). Nowadays, this field is widespread and used by most of the scientific community. Its interest does not need to be discussed further.
- *Graph theory* Graph theory is the study of graphs, which are mathematical structures used to model pairwise relations between objects. A graph contains vertices, or nodes, representing objects, as well as edges, or links, which depict relationships between nodes (Newman 2003; Diestel 2000). As an example, graphs can be used to represent networks of publications with authors and institutions as nodes and their respective relationships as edges.
- *Data mining* Data mining (Fayyad et al. 1996) is the process of discovering hidden information (called knowledge) and meaningful structures from large databases. It uses both supervised and unsupervised learning algorithms to cluster, classify, explain and predict data. Specifically, it can help to discover, describe and predict links or trends within data.
- *OLAP analysis* OLAP (Online Analytical Processing) (Chaudhuri and Dayal 1997) is the technology that exploits information in data warehouses. OLAP allows a multidimensional data analysis by building cubes. Through these cubes, it provides easy navigation, visualization and fast analysis for decision making within vast amounts of data.

Approaches that deal with bibliographic data

In Table 1, we present some research works according to their goal(s) and to the type of analysis, that is to say the field(s) they are coming from.

Hudomalj and Vidmar (2003), Baid et al. (2008) and Trifonova (2011) are interested in bibliographic data warehousing and OLAP processing systems because they are suitable for performing complex queries on large datasets. These authors implemented OLAP systems that include multidimensional exploration and, often, search engines. Users can interactively browse hierarchical summarized data. For example, the `Biomedicina Slovenica OLAP` system provided relevant biomedical and life sciences data (Hudomalj and Vidmar 2003). It used a star schema to model data extracted from the Slovenian national bibliographic database, which covers biomedical and life sciences publications. The authors implemented an OLAP solution to promptly analyse Slovenian bibliographic data. The system provided different results, such as the evolution over time of the amount of published papers, of citations; the papers' dependence on the number of co-authors and the number of organizations the authors are affiliated to, etc. Another example is the `DBPubs` prototype which integrated keyword search and OLAP operations to analize and explore the publications' content (Baid et al. 2008). Due to the query result containing thousands of papers, and in order to discover trends and to rank authors, this system applied a statistical analysis by computing scores for papers based on links between

Table 1 Several approaches dealing with bibliographic data

References	Types of analysis				Research goals						
	OLAP	Graph theory	Data mining	Statistics	Search engine	Relationship	Ranking	Community mining	Topic detection	Multidim exploration	Prediction
Trifonova (2011)	X				X					X	
Hudomalj and Vidmar (2003)	X				X					X	
Baid et al. (2008)	X			X	X	X	X			X	
Klink et al. (2004)		X	X		X	X					
Klink et al. (2006)		X		X	X	X					
Zaiane et al. (2009)		X	X		X	X		X	X		
Muhlenbach and Lallich (2010)		X	X		X			X			
Pham and Klamma (2010)		X	X			X					
Varlamis and Tsatsaronis (2011)		X	X			X		X			
Coscia et al. (2009)		X	X	X				X			
Deng et al. (2008)		X	X	X			X		X		
Huang et al. (2009)			X	X		X		X			X
Seki et al. (2010)				X			X				
Cabanac (2011)			X	X	X	X					

documents (such as citations). Then, it aggregated papers to find the seminal ones. The bgMath/OLAP system, also developed in the scientific literature, aimed to allow an easy manipulation while monitoring, evaluating and comparing scientific fields (Trifonova 2011).

Graphs are often used to visualize relationships between data, relationships that are not apparent while searching and browsing data. Concerning bibliographic data, graphs are currently used to show relationships between conferences and journals or authors. Klink et al. (2004, 2006) proposed DBLBrowser, a user friendly interface for searching, browsing, and mining bibliographic data. Their system combined both textual and visual browsing functionalities. It could find related publications and their correct bibliographic data. During the browsing process, data is visualized by suitable graphical techniques which help users to understand their research domain, to find relevant authors or publications, and above all, to provide information about distant researchers and relevant conferences or journals. Zaiane et al. (2009) introduced DBconnect, a prototype that performs social network analysis in the DBLP database. They rely on a new random walk approach to reveal interesting knowledge about the research community and even to recommend collaborations. The system looked for pertinent research communities, relevant conferences, similar authors, interesting topics, etc. It combined a random walk algorithm, text mining techniques and social network analysis to compute relevance scores between data and then extract information. Muhlenbach and Lallich (2010) proposed a matrix formalization that considers the similarity and dissimilarity between social relationships. They tried to discover research communities with a clustering method by using the neighborhood graph obtained with the dissimilarity scoring. A graph-theoretic model to locate research communities within the DBLP database is also introduced. Pham and Klamma (2010) clustered research communities from similar venues. They were interested in the structure of networks in Computer Science journals, conferences and workshops using citation analysis. Social network analysis (SNA) was applied to determine clusters of venues by calculating two network analysis measures for each one of them: betweenness and PageRank. Varlamis and Tsatsaronis (2011) proposed a new bibliographic data model to identify future research from a co-authorship network. The new representation model combines co-authorship and content similarity information. The authors used a graph visualization tool from the biological domain to provide comprehensive visualizations that help users to uncover hidden relations between authors. It also suggests potential synergies between researchers or groups. Gupta et al. (2011) considered the two problems of clustering and evolution diagnosis in bibliographic networks. They presented an algorithm, *ENetClus*, which performs, with a temporal smoothness approach, agglomerative evolutionary clustering that is able to show variations in the clusters over time. They calculated a probabilistic generative model from each cluster. Next, they evaluated objects in each cluster with a maximum likelihood approach, including a ranking condition in current and previous clusters.

Many researches have proposed a framework which combined data mining and statistics to deal with bibliographic data. Coscia et al. (2009) defined the problem of analyzing bibliographic data as the analysis of a data set of graphs, instead of a unique large graph. They adopted existing data mining and graph mining techniques to validate, compare and enrich diverse statistical parameters. They also used co-clustering (i.e., simultaneous partitioning of rows and columns of a contingency matrix) to assign a research profile (based on frequently used keywords) to each author. Deng et al. (2008) proposed three models to find experts based on large bibliographic databases. First, with a novel weighted language model, they found an expert candidate based on the relevance and importance of

its associated documents by introducing a document prior probability. The second is a topic-based model. It represents each candidate as a weighted sum of multiple topics. The third, a hybrid model, combines the language model and the topic-based model. Huang et al. (2009) aimed to detect the evolution of semantic communities. Based on keywords, they clustered the communities into two categories: giant community and small community. They created a word-association network based on keywords' co-occurrence in titles. More specifically, they analyzed the distribution of the edges frequency (also known as degree distribution). Seki et al. (2010) studied the current and future impact of social bookmarks on bibliographic information systems. They tried to compare social tags with conventional ones to improve the information retrieval performance. In the context of scientific literature recommendation systems, which suggests papers to researchers according to their scientific interests, Cabanac (2011) processed publications' *freely available metedata*. He designed a new inter-researcher similarity measure based on *topical and social clues* as they reflected the proximity and the strength of a relationship between researchers.

Business Intelligence being our field of research, we are more interested in the online analysis of bibliographic data.

OLAP on information networks

In this section, we explain the basic concepts of information networks and traditional OLAP analysis on classical data. Nevertheless, OLAP must change if we want to make online analysis of data from information networks that is usually modeled as graphs. At the end of the section, we propose a comparison between traditional OLAP and OLAP on information networks (or Graph OLAP).

Information networks

An information network is made of numerous interacting and multi-typed objects. Graphs have been widely used to model networks. A graph $G = (V, E)$ consists of V, a set of vertices (or nodes) and E, a set of edges (or links). Each edge has two vertices associated with it. A node can be connected by one or more links. Each node represents an object or an entity. An edge represents a relationship between two nodes. For example, in bibliographic networks, entities can be authors, publications, institutions, conferences, etc. Links may depict author, co-author, *belongs to* or other kinds of relationships. They may also include a label or a weight. Apart from the topological structure encoded in the underlying graph, multidimensional attributes are often specified and associated with vertices, forming the so-called multidimensional networks (Zhao et al. 2011). A multidimensional network is defined as a graph $G = (V, E, A)$, where A is a set of n vertex-specific attributes. A is called the dimensions of the network. In the co-authorship network (Fig. 1a), each node represents an author and the associated attributes can be gender, age, etc.

There are two types of networks. In the first type, networks are homogeneous. They contain a single object type and a single link type such as co-authorship networks. The co-authorship network (or the authors network) is a homogeneous network: each node represents an author; each edge between two authors represents a co-author relationship, in one or several papers, with attributes like conference, year and venue (Fig. 1a). There may be multiple edges between two nodes if two authors have co-written more than one paper

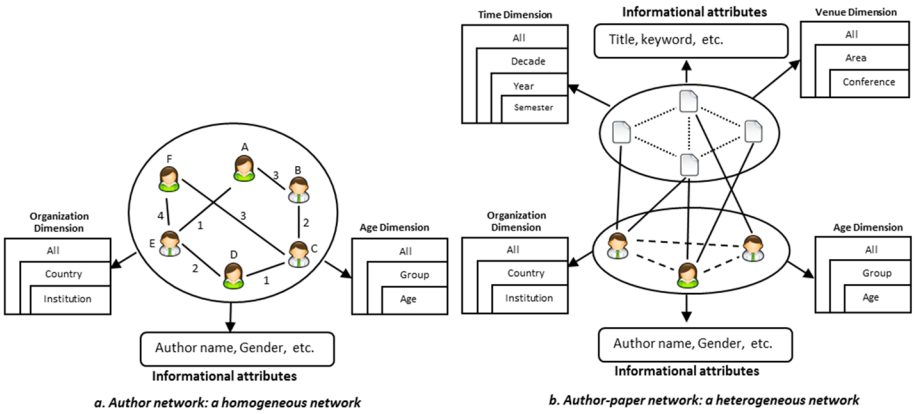

Fig. 1 Examples of bibliographic networks

together. For instance, authors A and B wrote together one paper in 2008 at ASONAM conference, one in CIKM 2009 and one in DASFAA 2010. In consequence, the weight 3 has been added over the edge between them. It means that author A and B have written three papers together. In the second type, networks are composed of multiple objects and link types. They are called heterogeneous networks. An example is given by the author-paper network (Fig. 1b). This network has two types of nodes: authors and papers. There are three types of edges. The first link is *written by* between authors and papers. The second represents co-author relationship and the last one relates papers written by the same authors. Each object is associated with a set of multidimensional attributes that describes it. For instance, the paper object has *venue* and *time* as attributes, as well as being also associated to *title* and *keywords*.

The concepts of homogeneous and heterogeneous networks are a generalization of those of one-mode networks (e.g. authors network with authors as nodes and co-authorship as links) and those of multi-mode networks (e.g. affiliations or memberships networks with one set of nodes like authors and multiple sets of links like co-authors or authors from the same conference, authors within similar conferences, authors with similar publications, etc.) introduced by Klink et al. (2006). They are also a generalization of bipartite (authors–conferences) and tripartite (authors–conferences–topics) graph models used for example by Zaiane et al. (2009).

Online analytical processing (OLAP)

OLAP analysis is a multidimensional data analysis used to provide fast interpretation for decision making within a large amount of data (Chaudhuri and Dayal 1997) providing easy navigation and visualization. OLAP gives a multidimensional view of data by building cubes. The multidimensional model consists of facts represented by measures and dimensions. The cube contains cells with measures, which are values based on a set of dimensions. Dimensions can be seen as an analysis axis and may be organized into hierarchies with several levels. For instance, in Fig. 2, the publications are the facts. The hierarchy of the venue dimension has three levels: the support (the name of the conference, of the journal, or of the book), the research area (databases, data mining, information retrieval, etc.) and the all level.

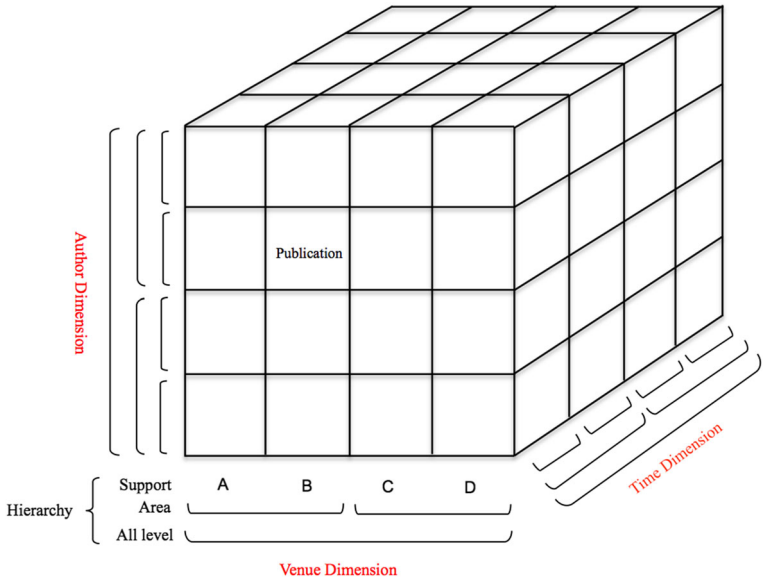

Fig. 2 Example of an OLAP cube

Essentially, measures can be numerical indicators which can be aggregated. An interesting feature of the multidimensional model is the measure aggregation according to one or more dimensions. For example, it is possible to compute the total number of publications by area over the years.

There are four classic OLAP operations: *roll-up* takes the current data and does a group-by on one dimension in order to aggregate or summarize facts; *drill-down* is the complement of the *roll-up* operator by giving more details; *slice and dice* reduce dimensions by taking a subset of them from the data and, finally, *pivot* changes the layouts according to different points of view.

A special feature of bibliographic data is that it can be seen as an information network. It is possible to build several networks such as a co-authors network, a citations network, a conferences network, etc. The goal of information network analysis is to understand the structure and the behavior of a given network. Extracting knowledge from large networks is a complex task and it is too big to be human-comprehensible. OLAP analysis could be a good approach in order to have a more compact view of the data.

Comparison between traditional OLAP and Graph OLAP

In the literature, there are several expressions for OLAP on heterogeneous information networks, also known as Graph OLAP. These expressions are a generalization of Social OLAP, which is OLAP on data coming from social networks. As we have explained before, OLAP must change if we want to perform online analysis over information networks data which is generally modeled as graphs. Consequently, we propose a comparison between traditional OLAP and what Graph OLAP is or should be. Our comparison is summarized in Table 2.

Traditionally, data warehouses are used to store, to model, to analyze and to visualize relational structured or semi-structured data, and more recently, textual and XML data.

Table 2 Comparison between traditional OLAP and Graph OLAP

	Traditional OLAP	Graph OLAP
Data	Relational or semi-structured data textual or XML data	Interconnected objects of different types
Problems	Not considering links among data records	How taking interactions among entities into account?
Input	Multidimensional facts	Networks with entities, links and attributes
Output	Aggregated measures	A new network more generalized
Dimensions	Only informational	Informational and topological
Hierarchies	Yes (informational attributes)	Yes (both for info. and topo. dimensions)
Measures	Numeric indicators	Measures coming from graph theory
	Aggregation function (count, sum, average)	Aggregated graph measure
		Specific aggregation functions
Operations	Informational operations: Roll-up, drill-down, etc.	Informational and topological OLAP operations

Data warehouses present information in tables with rows and columns. A table is a collection of objects (records or rows) of the same type. Relationships occur between tables yet records are not considered as interconnected (or interrelated) objects. However, in Graph OLAP, information is interconnected and it is in the form of networks. In real applications, networks contain several complex types of relationships. It is difficult to explore information in-depth when there exists numerous relationships. We believe that heterogeneous information networks can be considered as a generalization of databases, as semi-structured data and even as a kind of corpus of documents. For example, from a database of publications such as DBLP or PubMed, where publications are linked via authors, citations, institutions, topics, etc., we can build a network of co-authors, a network of citations, a network of conferences, etc.

In traditional OLAP, cubes contain facts defined by dimensions and measures. Aggregates are obtained with the use of operators like *roll up*. In this context, aggregates are facts whose measures were synthesized according to certain dimensions. In graph OLAP, cubes can contain graphs as input. Graphs are defined by a structure (entities and edges) and attributes. The aggregation of a graph gives a more general graph as output. For example, the whole author-paper network (Fig. 3a) could be too big to be comprehensible, and thus it could be a good idea to look at it from a more compressed view. One may want to see the authors collaborations according to their institutions. This task requires the network to be generalized by merging all authors of the same institution as one node and building a new summary graph at the institution level (Fig. 3b). In this more generalized network, an edge between Stanford and the university of Lyon will aggregate all the collaborations occurred between Stanford's authors and the authors from the university of Lyon. The difference with a classical *roll up* is that a *roll up* from the individual level to the institution level is achieved by consolidating multiple nodes into one, which in turn shrinks the whole graph.

Using Chen's words, in Graph OLAP, there may be several types of dimensions: informational dimensions (as in traditional OLAP) and topological dimensions (Chen et al. 2008). In the author–paper network, *venue* and *time* are two informational attributes. They can be used as informational dimensions with their respective hierarchies: *semester, year, decade, all*; and *support, research area, all*, respectively. For example, these attributes

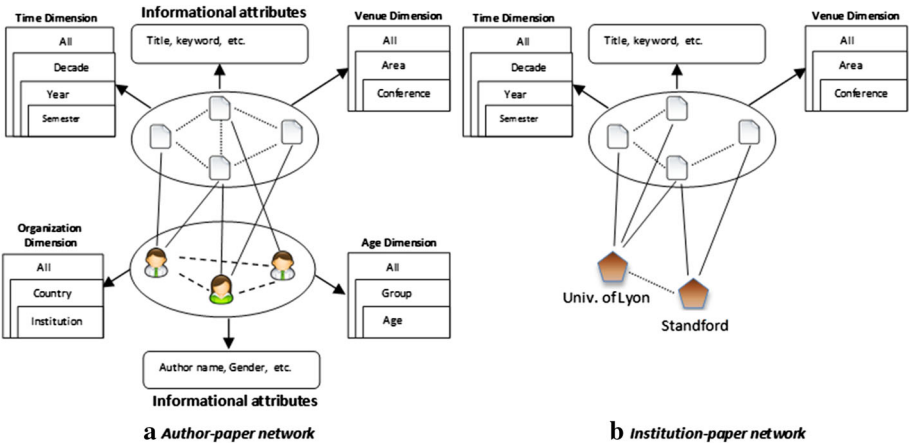

Fig. 3 Example of aggregated network

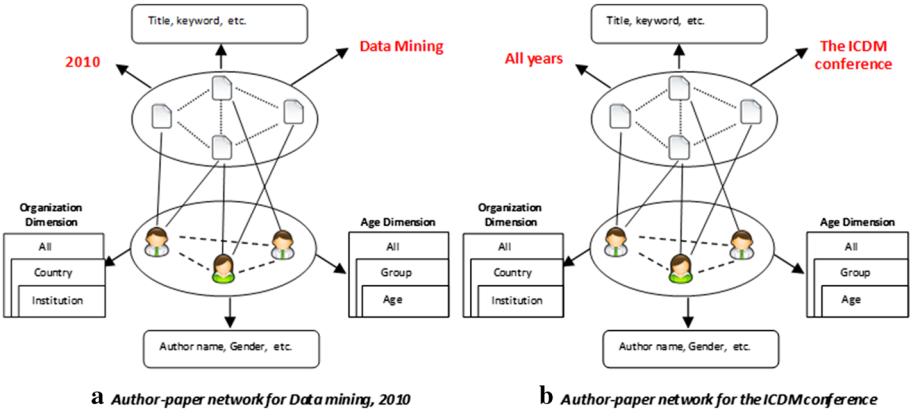

Fig. 4 Example of networks

allow building a network of authors for the ICDM conference over all the possible years (Fig. 4a), as well as a network for the data mining field in 2010 (Fig. 4b). The organization dimension, with its hierarchy *institution, country, all*, can be used as a topological dimension and allows merging all authors from the same institution in a more general node (Fig. 3b). A new graph with more generalized nodes is created by summarizing the original network. Within networks, topological dimensions operate on nodes and edges. We think that topological dimensions are a real added value because they allow us to model the relationships between objects.

Measures are traditionally numeric (like indicators). They are associated with an aggregation function, such as sum, average or count, in order to synthesize and aggregate facts. In an XML, or text, data warehouse we can find other types of measures with the appropriate aggregation functions. In Graph OLAP we can have classical measures like the number of publications, and also numeric measures coming from graph theory (closeness, centrality degree, diameter, etc.). However, the measure can also be a graph. A graph can

be both a fact and a measure, which goes against the classical principles of multidimensional modeling. If the measure is a graph, then it is necessary to develop new aggregation functions adapted to it. We believe that the aggregation of a graph should take into account both the attributes describing the entities and the relationships between entities.

In traditional OLAP, when a *roll up* is made on an informational dimension, the network structure does not change, as this is an informational OLAP operation. In contrast, a *roll up* on a topological dimension reorganizes the network for a more generalized view. It is then a topological OLAP operation. The topological structure of the original graph is modified. Chen et al. (2008) speak about *I-OLAP* and *T-OLAP* operations.

In the end, traditional OLAP is a special case of Graph OLAP when we are not taking into account the relationships between facts.

Literature review

The topic of OLAP on information networks is quite new. Only few research teams have been interested in this topic. To the best of our knowledge, the first works were published around the year 2008.

Han's team and his colleagues were among the first to investigate OLAP on information networks (Chen et al. 2008; Qu et al. 2011; Jin et al. 2010; Zhao et al. 2011). Chen et al. (2008) presented the basic definitions of OLAP on information networks and introduced a general framework called `Graph OLAP`. Qu et al. (2011) worked on topological OLAP operations to allow *roll-up* procedures on topological dimensions by changing the structure of the aggregated graph. The key problem is to efficiently compute measures for the newly aggregated networks and to handle user queries with various constraints. Two effective computational techniques, `T-Distributiveness` and `T-Monotonicity` were proposed to achieve efficient query processing and cube materialization. Zhao et al. (2011) defined the concept of multidimensional networks to abstract real networks. They also introduced a new multidimensional model, called `Graph Cube`, to extend data warehouses to large multidimensional networks. They worked with structure-enriched aggregate networks and proposed a new type of query for multidimensional networks, called `crossboid query`. In contrast with traditional queries, known as `cuboid queries`, a `crossboid query` can cross more than one cube in a query rather than a single cube, as in `cuboid queries`. The `Graph Cube` model also considers aggregation networks both on entities and relationships. Jin et al. (2010) proposed the `Visual Cube` model as well as OLAP analysis for image collections, such as web images indexed in search engines, product images or photos shared on social networks. `Visual Cube` provided online answers to user requests with summarized statistics of image information and helped users navigate and explore images efficiently. Four measures have been presented in the `Visual Cube` model. The first measure summarizes information as in traditional OLAP. The other measures are unique for `Visual Cube`: summarized image feature (i.e. average color histogram), subset of images (i.e. clustering images and choosing the central one), and all images (ranking lists).

With regard to summarizing attributed networks in the context of OLAP analysis, the closest works to those of Han's team are those of Tian et al. They introduced two operations to summarize graphs in OLAP analysis (Tian et al. 2008). The first operation, named `SNAP` (Summarization by Grouping Nodes on Attributes and Pairwise Relationships) merges homogeneous nodes, combines corresponding edges and aggregates a graph

that displays relationships for generalized nodes. The second one, called k-SNAP, allows users to control the size of summarized graphs by specifying the number of k groups.

Morfonios and Koutrika (2008) researched social bookmaking systems and they were also pioneers in the field of OLAP on information networks. They proposed going beyond classical keyword-based searching to exploring social data starting from any type of entity (user, resource or annotation) and requesting aggregated views of related entities based on the relationships defined between entities. Then, they mapped this type of social searching to OLAP query processing and they studied various ways to support on-the-fly aggregations of data. Finally, they described how data cubes can be used to precompute and materialize the results of all possible aggregate queries over social data. Similarly, Wu et al. (2012) worked with user profiles on social networks. They proposed an OLAP serving system, called Avatara, to handle many small cubes. The system provides a simple, expressive grammar for application developers to construct cubes and query them at scale.

Yin et al. (2012) criticized Chen's model ability to handle only homogeneous networks. They defined the concept of entity dimensions to complement informational and topological dimensions and to handle heterogeneous networks. They also introduced two OLAP operations: Rotate, to convert entities into relations and vice versa; and Stretch, to discover implicit relationships between entities. Their third contribution consists in two new models: HMGraph OLAP, a new multidimensional model of data warehouse for heterogeneous networks; and HMGraph Cube, a model for aggregating cubes of graphs.

Beheshti et al. (2012) disapproved the existing approaches for supporting only multidimensional and multi-level queries on graphs, for not providing a semantic-driven framework and for not supporting a language for n-dimensional computations. N-dimensional computations are frequent in OLAP analysis. For example, it could be interesting to analyze the reputation of a book, an author, or a publisher in a specific year. Such a query requires supporting n-dimensional computations on graphs, providing multiple views at different granularity levels. Consequently, the authors proposed a graph data model, called GOLAP that extended decision support on multidimensional networks and considered both objects and links. They used the concepts of folder and path nodes to support multidimensional and multi-level views and to provide network semantics. Traditional dimensions and measures are redefined according to the relationships among entities. Finally, they also extended the SPARQL language in order to support n-dimensional computations on graphs and proposed new OLAP operations (assignment, function, update, upsert).

Kampgen and Harth (2011) retrieved statistical information from multiple linked data sources to insert them into a data warehouse. The authors proposed a mapping between linked data and multidimensional models by using the RDF Data Cube vocabulary in order to take into account data semantics. It is regrettable that the mapping is relatively conventional with only traditional OLAP concepts and without taking into account the topological structure of the networks.

Kaya and Alhajj (2014) joined two databases, DBLP and CiteSeerX, in order to have bibliographic information on major computer science conference proceedings and journals and to include citations, co-authorships, addresses, and affiliations from authors. They developed three different information networks (Authors, Topic and Venue) with a cube-based modeling method. In these networks, each node may represent an author, a topic or a venue respectively. Next, each node is represented by a data cube which is later analyzed by OLAP. Finally, using a multi-agent based algorithm, they automatically found relevant persons, topics and venues for each network respectively.

Table 3 Works about OLAP on information networks

Paper	Data	Network	Measures	Aggregation function	Dimensions			Materialization			Operations
					I	T	Te	F	P	N	
Graph OLAP (Chen et al. 2008)	Bibliographic data	Homogeneous	Coming from	Aggregated graph	X	X		X	X		Topological
Graph Cube (Zhao et al. 2011)	Bibliographic data	Homogeneous	Graph theory	Aggregated graph	X	X		X	X		Topological
Qu et al. (2011)	Bibliographic data	Homogeneous	Or a graph	Aggregated graph	X	X			X		Topological
Visual Cube (Jin et al. 2010)	Images	Homogeneous	Image or image features	Clustering aggregation	X						Classical
Tian et al. (2008)	Bibliographic data	Homogeneous	Graph	Aggregated graph	X	X					SNAP k-SNAP
Morfonios and Koutrika (2008)	Social networks	Heterogeneous	Number of relations	COUNT	X			X			Classical
Avatar (Wu et al. 2012)	Social networks	Homogeneous	Classical	Classical	X						Classical
Yin et al. (2012) HM Graph OLAP HM Graph Cube	Bibliographic data	Heterogeneous	Graph	Aggregated graph	X	X	X		X	X	Rotate Stretch
GOLAP (Beheshti et al. 2012)	Bibliographic data	Heterogeneous	Graph	Aggregated graph	X	X					Assignment Update Upsert
Kampgen and Harth (2011)	Statistical data		Classical	Classical	X						Classical
Kaya and Alhajj (2014)	Bibliographic data	Homogeneous	Classical	Classical	X						Classical

Discussion

To conclude this survey paper and to sum up the work related to OLAP on information networks, we propose a comparison between the approaches in Table 3. The two first criteria in the table recall the data or domains which are studied and the type of networks built from these data (homogeneous or heterogeneous). The other criteria deal with how information networks are designed in the multidimensional model and also show how the works adapt OLAP to networks. For each approach, the type of measure and the associated aggregation function are indicated. There can be several kinds of dimensions: informational dimension (I), topological dimension(T) and entity dimension (Te). Some works focus on efficient computation of cubes and users' queries and propose a full materialization (F), a partial materialization(P) and a non materialization (N). Finally, some specific OLAP tools or operations are sometimes created to answer users' queries.

Most approaches deal with bibliographic data because it is well known and constitute a suitable example of information networks. Usually, a co-authors network is built and it has different attributes such as time, venue, area and so on. Zhao et al. added an attribute, namely productivity, by discretizing the publications number of an author into four different buckets (Excellent, good, fair and poor). Sometimes, approaches deal with other kinds of data such as images (Jin et al. 2010), social networks (Morfonios and Koutrika 2008; Wu et al. 2012) and statistical data (Kampgen and Harth 2011). Regarding preprocessing, Kampgen et al. are the only team to mention an ETL process for extracting, transforming and loading linked data into a data warehouse.

The two main limits of the studies (Chen et al. 2008; Zhao et al. 2011; Qu et al. 2011; Jin et al. 2010; Tian et al. 2008; Wu et al. 2012) are that only homogeneous networks are built and usually only one network. We think it would be better to build heterogeneous networks as proposed in Morfonios and Koutrika (2008), Yin et al. (2012), Beheshti et al. (2012) and to build, from the same database, several networks (some of them being heterogeneous) in order to take into account multiple points of view. Studying co-authors, citations, topics and conferences networks could give different points of view from the same database. Nonetheless, to the best of our knowledge, no approach actually does this.

The multidimensional model for networks is quite different from the traditional model. There is a redefinition of dimensions, measures and operations in order to adapt them to graphs and networks. As we said before, Han's team was the first to investigate OLAP on information networks. They introduced basic definitions through their general framework, called `Graph OLAP` (Chen et al. 2008). The `Graph OLAP` framework was formally used by other research teams. We have always found the same concepts definitions for topological and informational dimensions, specific measures and aggregation functions. Most of the time, the measure is a graph or comes from graph theory such as centrality degree (Qu et al. 2011), number of relations (Morfonios and Koutrika 2008), etc. When the measure is a graph, all approaches defined an aggregation function adapted to graphs. We think that the model must take into account multiple types of measures and not only one. For each type of measure, there should be an adapted aggregation function. For example, if the measure is a centrality degree, how can it be aggregated when a *roll-up* is done? The aggregation function of a graph should also take both entities and structure into account. Another example is clustering entities into groups that share similar properties. Then it is possible to have an aggregation function like that of Jin et al. (2010).

Only one approach, that of Yin et al. (2012), augmented the dimensions with the concept of entity dimension, which aims to handle heterogeneous networks. They also

included two fact tables in the multidimensional model: one for entities and one for relationships between entities.

With the introduction of topological dimensions, authors introduced topological OLAP operations. Furthermore, Tian et al. (2008) proposed new operations for summarizing graphs. Users can freely choose the interesting attributes and relationships. In contrast, Yin et al. (2012) and Beheshti et al. (2012) proposed new operations to view knowledge inside graph cubes. We believe that to visualize networks, to extract knowledge from them, to analyze their dynamics (such as most popular topics over time, etc.) new OLAP tools must be created by combining data mining, social networks analysis, information retrieval and OLAP philosophies. We think it is a promising research issue.

Finally, thanks to the literature review and the comparison of traditional OLAP and Graph OLAP, we have identified and discussed several challenges while dealing with bibliographic data using OLAP and information networks. The first challenge is to build a data warehouse for several heterogeneous networks. The second challenge is to design a multidimensional model for multiple heterogeneous networks. We think that classical models cannot meet our needs and we are probably compelled to create a new model. Lastly, there is the crucial challenge of providing analysis tools able to handle the diverse networks considered. Innovative tools should be developed for users. We plan to combine data mining or text mining methods, information retrieval approaches and social networks analysis with OLAP operators. As publications contain textual data, this type of information must be extracted, represented, and analyzed. The introduction of text mining techniques into the workflow of bibliographic analysis is already a promising research direction.

References

Baid, A., Balmin, A., Hwang, H., Nijkamp, E., Rao, J., Reinwald, B., Simitsis, A., Sismanis, Y., & Ham, F. (2008). Dbpubs: Multidimensional exploration of database publications. In *Proceedings of the 34th international conference on very large data bases* (Vol. 1, pp. 1456–1459).

Beheshti, S., Benatallah, B., & Motahari-Nezhad, H. (2012). A framework and a language for on-line analytical processing on graphs. In *13th international conference on web information systems engineering (WISE'12)* (pp. 213–227).

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, *87*(3), 597–620.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and olap technology. *ACM SIGMOD*, *26*(1), 65–74.

Chen, C., Yan, X., Zhu, F., Han, J., & Yu, P. (2008). Graph olap: Towards online analytical processing on graphs. In *IEEE international conference on data mining (TCDM'08)* (pp. 103–112).

Coscia, M., Giannotti, F., & Pensa, R. (2009). Social network analysis as knowledge discovery process: a case study on digital bibliography. In *Proceedings of the 2009 international conference on advances in social networks analysis and mining (ASONAM '09)* (pp. 279–283).

Deng, H., King, I., & Lyu, M. (2008). Formal models for expert finding on dblp bibliography data. In *Proceedings of the 2008 eighth IEEE international conference on data mining (ICDM'08)* (pp. 163–172).

Diestel, R. (2000). *Graph theory* (2nd ed.). New York: Springer.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD Proceedings* (pp. 82–88).

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics*, *93*(3), 765–785.

Gupta, M., Aggarwal, C., Han, J., & Sun, Y. (2011). Evolutionary clustering and analysis of bibliographic networks. In *International conference on advances in social networks analysis and mining (ASONAM'11)* (pp. 63–70).

Huang, Z., Yan, Y., Qiu, Y., & Qiao, S. (2009). Exploring emergent semantic communities from dblp bibliography database. In *International conference on advances in social network analysis and mining (ASONAM'09)* (pp. 219–224).

Hudomalj, E., & Vidmar, G. (2003). Olap and bibliographic databases. *Scientometrics*, *58*(3), 609–622.

Hulme, E. W. (1923). *Statistical bibliography in relation to the growth of modern civilization*. London: Grafton.

Jakawat, W., Favre, C., & Loudcher, S. (2013). Olap on information networks: A new framework for dealing with bibliographic data. In *1st International Workshop on Social Business Intelligence (SoBI 2013), collocated with the East-European Conference on Advances in Databases and Information Systems (ADBIS)* (pp. 361–370).

Jin, X., Han, J., Cao, L., Luo, J., Ding, B., & Lin, C.X. (2010). Visual cube and on-line analytical processing of images. In *19th ACM international conference on Information and knowledge management (CIKM'10)*.

Kampgen, B., & Harth, A. (2011). Transforming statistical linked data for use in olap systems. In *7th international conference on semantic systems (I-SEMANTICS'11)* (pp. 33–40).

Kaya, M., & Alhajj, R. (2014). Development of multidimensional academic information networks with a novel data cube based modeling method. *Information Sciences*, *265*, 211–224.

Klink, S., Ley, M., Rabbidge, E., Reuther, P., Walter, B., & Weber, A. (2004). Visualising and mining digital bibliographic data. In *INFORMATIK* (pp. 193–197).

Klink, S., Reuther, P., Weber, A., Walter, B., & Ley, M. (2006). Analysing social networks within bibliographical data. In *Proceedings of the 17th international conference on database and expert systems applications (DEXA'06)* (pp. 234–243).

Morfonios, K., & Koutrika, G. (2008). Olap cubes for social searches: Standing on the shoulders of giants? In *International workshop on the web and databases (WebDB)*.

Muhlenbach, F., & Lallich, S. (2010). Discovering research communities by clustering bibliographical data. In *IEEE WIC ACM international conference on web intelligence and intelligent agent technology (WI-IAT'10)* (Vol. 1, pp. 500–507).

Newman, M. E. (2003). The structure and function of complex networks. SIAM review, *45*(2), 167–256.

Pham, M.C., & Klamma, R. (2010). The structure of the computer science knowledge network. In *International conference on advances in social networks analysis and mining (ASONAM'10)* (pp. 17–24).

Pritchard, A. (1969). *Statistical bibliography: An interim bibliography*. London: North-Western Polytechnic, School of Librarianship.

Qu, Q., Zhu, F., Yan, X., Han, J., Yu, P., & Li, H. (2011). Efficient topological olap on information networks. In *Proceedings of the 16th international conference on database systems for advanced applications (DASFAA'11)* (Vol. 1, pp. 389–403).

Seki K., Qin, H., & Uehara, K. (2010). Impact and prospect of social bookmarks for bibliographic information retrieval. In *Proceedings of the 10th annual joint conference on digital libraries (JCDL '10)* (pp. 357–360).

Tian, Y., Hankins, R., & Patel, L. (2008). Efficient aggregation for graph summarization. In *ACM SIGMOD international conference on management of data (SIGMOD'08)* (pp. 567–580).

Trifonova, T. G. (2011). Warehousing and olap analysis of bibliographic data. *Intelligent Information Management*, *3*, 190–197.

Van Raan, A. F. J. (1997). Scientometrics: State-of-the-art. *Scientometrics*, *38*(1), 205–218.

Varlamis, I., & Tsatsaronis, G. (2011). Visualizing bibliographic databases as graphs and mining potential research synergies. In *Proceedings of the 2011 international conference on advances in social networks analysis and mining (ASONAM '11)* (pp. 53–60).

Wu, L., Sumbaly, R., Riccomini, C., Koo, G., Kim, H., Kreps, J., et al. (2012). Avatara: Olap for webscale analytics products. *Proceedings of the VLDB Endowment*, *5*(12), 1874–1877.

Yin, M., Wu, B., & Zeng, Z. (2012). Hmgraph olap: a novel framework for multi-dimensional heterogeneous network analysis. In *15th international workshop on data warehousing and OLAP (DOLAP'12)* (pp. 137–144).

Zaiane, O. R., Chen, J., & Goebel, R. (2009). Mining research communities in bibliographical data. *Advances in Web Mining and Web Usage Analysis*, *5439*, 59–76.

Zhao, P., Li, X., Xin, D., & Han, J. (2011). Graph cube: On warehousing and olap multidimensional networks. In *ACM SIGMOD international conference on management of data (SIGMOD'11)* (pp. 853–864).