iNextCube: Information Network-Enhanced Text Cube *

Yintao Yu Cindy X. Lin Yizhou Sun Chen Chen Jiawei Han
Binbin Liao Tianyi Wu ChengXiang Zhai Duo Zhang Bo Zhao
Department of Computer Science
University of Illinois at Urbana-Champaign
{yintao, xidelin2, sun22, cchen37, hanj, liao17, twu5, czhai, dzhang22, bozhao3}@uiuc.edu

ABSTRACT

Nowadays, most business, administration, and/or scientific databases contain both structured attributes and text attributes. We call a database that consists of both multidimensional structured data and narrative text data as *multidimensional text database*. Searching, OLAP, and mining such databases pose many research challenges. To enhance the power of data analysis, interesting entities and relationships can be extracted from such databases to derive heterogeneous information networks, which in turn will substantially increase the power and flexibility of data exploration in such databases. Based on our previous studies on TextCube [1], TopicCube [2], and information network analysis, such as RankClus [3] and NetClus [4], we construct iNextCube, an **i**nformation-**N**etwork-**e**nhanced te**xt Cube**. In this demo, we show the power of iNextCube in the search and analysis of two multidimensional text databases: (i) a DBLP-based CS bibliographic database, and (ii) an online news database.

1. INTRODUCTION

With the boom of Internet and the ever increasing business intelligence applications, search and analysis of text data have attracted broad attention. Among a large variety of text data, we pay special attention to one particular kind of text data, *multidimensional text database*, which is the database that consists of both multidimensional structured data and narrative text data. Most business, adminstration, and/or scientific databases also contain narrative attributes besides structured data and thus belong to this category. Searching, online analytical processing (OLAP), and mining of such text databases pose many research challenges. In this demo proposal, we focus on search and analysis of such databases and propose to demonstrate a novel, integrated text OLAP system: iNextCube, *i.e.*, **i**nformation-**N**etwork-**e**nhanced te**xt Cube**.

Traditional OLAP cube studies have been focused on structured data with numeric measures. Recent years have seen OLAP cubes extended to new domains, such as graphs, sequences, XML data, spatial data and mobile data. There are many existing methods to summarize text databases. However, although some previous studies have proposed OLAP cube systems, none of them provides a general cube model to meet the most urgent and challenging demand—to perform advanced analysis of unstructured text information along with structured categorical information in such a way that these two kinds of information can mutually enhance knowledge discovery and data analysis. Our recent studies on TextCube [1] and TopicCube [2] propose such an OLAP cube model on multidimensional text database that summarizes and navigates structured data together with narrative text data for integrated, OLAP-based IR applications.

This method, though interesting and powerful, encounters one knowledge acquisition bottleneck: it is highly desirable to have multi-level concept hierarchies on structured data dimensions and keyword/term dimensions in such databases, such as hierarchies of research subareas and term hierarchies for computer science in the DBLP data. However, it is time-consuming and error-prone to rely on domain experts to provide such knowledge. Our recent study on information network analysis proposed an interesting method, RankClus [3] and its extension NetClus [4], that helps construction of such hierarchies automatically by information network analysis.

To promote systematic design and development of scalable and effective methods for search, OLAP, and mining of multidimensional text database, we have been developing a research prototype system, iNextCube, that integrates our two recent research themes, text OLAP and information network analysis, into one Web-based system that accomplishes the following tasks: (i) *automated ranking, clustering, and hierarchy generation by* NetClus [4], (ii) *construction of* TextCube [1] *and* TopicCube[2] *in* iNextCube *system*, and (iii) *data mining in* iNextCube. The functionalities of the system will be demonstrated with two multidimensional text databases: (i) a computer science bibliographic database generated from DBLP, and (ii) a news text database extracted from the web. We select them for demo because they are typical multidimensional text databases, with different portions of narrative text data, and their entities and relationships can be

*The work was supported in part by the U.S. National Science Foundation grants IIS-08-42769 and BDI-05-15813, NASA grant NNX08AC35A, and the Air Force Office of Scientific Research MURI award FA9550-08-1-0265. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

VLDB '09, August 24-28, 2009, Lyon, France

extracted to form heterogeneous (*i.e.*, multi-typed) information networks. Moreover, they are interesting, understandable by the public, and are derived from massive, relatively complete, cleansed, and publicly available datasets.

2. GENERAL SYSTEM ARCHITECTURE

iNextCube has a four-layer architecture, as shown in Fig. 1. The lower intermediate layer is the NetClus [4] module that analyzes information networks and generates clusters, rankings, and concept hierarchies. The upper intermediate layer consists of the TextCube [1] and TopicCube [2] infrastructures, which provide online information retrieval measures and probabilistic latent semantic analysis (PLSA) [5] parameters. The top layer interacts with users and responds to their requests, including a web-based, user-friendly interface and visualization.

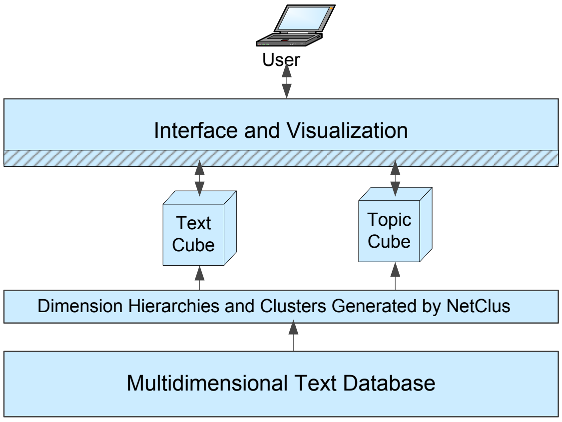

Figure 1: Architecture of iNextCube

3. MAJOR FUNCTIONAL MODULES

For lack of space, our illustration of the major functional modules is confined to the DBLP bibliographic database only. However, any multidimensional text database that consists of multidimensional structural attributes and narrative text attributes could be used as the data source.

3.1 Formation of dimensional hierarchies by information network analysis

It is tedious, error-prone, and non-scalable to rely on domain experts to specify concept hierarchies for cube dimensional data such as research areas, and for keywords and terms; and it is interesting to automatically construct concept hierarchies on these dimensions by information network analysis.

Our recent study on information network analysis has developed a RankClus method [3], by viewing the DBLP database as a *bi-type information network* with *conferences* as one type and *authors* as the other. By applying a few general rules derived from domain experts, *e.g.*, "highly ranked *authors publish* many *papers in* highly ranked *conferences*", and "highly ranked *conferences attract* many *papers from* many highly ranked *authors*", RankClus performs integrated clustering and ranking effectively and efficiently, and generates meaningful clusterings and rankings automatically without referring to any citation or content information in DBLP, with performance comparable to that of human experts. These clustering and ranking results are the basis of iNextCube for effective concept hierarchy formation, search, OLAP, and network mining.

In our system, NetClus, dealing with multi-typed (*i.e.*, ≥ 2 types) information network is used for integrated clustering, ranking, and concept hierarchy formation. A concept hierarchy is built for research areas using the information network extracted from the DBLP data, and a ranking distribution is generated for each type of entities at each node. The top level of the hierarchy, *i.e.*, the root, is the whole computer science area. All the entities including conferences, authors, papers and terms are 100% belong to that node. At the next level (*i.e.*, the first level), eight research areas are generated as children node of the root, which are "Database and Information Systems", "Hardware and Architecture", and so on. At deeper levels, for example, from the node "Database and Information Systems", there are further children nodes: "Database", "Data Mining" and "Information Retrieval"; and then from the node "Databases", its further children nodes are "Query Processing and Optimization", "XML", *etc.*. An example ranking result for each type of entities (authors/conferences/terms) in the node "Database" is shown in Table 1.

Based on this concept hierarchy and entity ranking scores in each node, the algorithm further clusters entities, *i.e.*, assigns entities into different nodes at each level. For example, for author "Serge Abiteboul", at the first level he belongs to "Database and Information Systems"; then at the next level, "Database"; and then at third level, "XML".

Notice that, for news dataset, entity extraction technology should be first applied. Accordingly, dimensions such as people, location, time, *etc.* will be formed and network containing these types of objects will be built through the process.

3.2 Search, OLAP, and Mining of Multidimensional Text Databases

A multidimensional text database contains both multidimensional structured data and unstructured text data. If we treat each paper as a record (a row), the *structured data* are the dimensions such as names of authors, conference, publication year, *etc.*, and the *unstructured text data* are the title or typically the abstract of the corresponding paper. On one hand, information network captures the structure information through the database; one the other hand, text information is accompanied with the objects in the information network.

Databases constructed from such information networks raise some new issues on search since the query contains both dimension information (user may want to specify a dimension, *e.g.*, conference) and text query (*e.g.*, "graph mining"). Although there are recent methods [6, 7, 8, 9] that perform content summary or OLAP on dimensional database, it is difficult to combine the power of both multidimensional OLAP and content summary. The reason is that without a support of some efficient infrastructure, it is difficult to respond to online queries efficiently for arbitrarily imposed query dimensions.

Recently, we have proposed two new infrastructures, called TextCube [1] and TopicCube [2] respectively, to support effective multidimensional text summary and search. In TextCube [1], TF (term frequency) and IDF (inverted document fre-

Conference	Rank Score
SIGMOD	0.315
VLDB	0.306
ICDE	0.194
PODS	0.109
EDBT	0.046
CIKM	0.019
...	...

Author	Rank Score
Surajit Chaudhuri	0.0065
Michael Stonebraker	0.0062
C. Mohan	0.0053
David J. DeWitt	0.0051
Jeffrey F. Naughton	0.0051
Michael J. Carey	0.0050
...	...

Term	Rank Score
database	0.0529
system	0.0322
query	0.0313
data	0.0251
object	0.0138
management	0.0113
...	...

Table 1: Ranking for each type of entities in "Database"

quency), which are two critical features to describe a document or a document set, are calculated efficiently in an online fashion for any given set of documents specified by the dimension values. In TopicCube [2], topic models for each set of documents specified by dimension values are generated according to a hierarchical prior tree from users' input. The two infrastructures make search, OLAP, and mining on multidimensional text databases possible.

3.2.1 Search

With the techniques of TextCube and TopicCube, queries including both structured dimension information and unstructured text information are able to be supported. We now examine three typical queries that users may submit to the system.

First, queries only contain dimension information. By submitting such queries, users may search for the text summaries and compare them among different dimension values. For example, users may compare the text summaries between paper set of conference "VLDB" in the year of "1998" and the same conference in the year of "2008". From TextCube and TopicCube, text summaries such as term frequency distribution and latent topic distribution can be effectively computed based on different settings of dimensions. Through text summarization for multiple dimension value combinations, users may obtain topic trends of a conference, topic differences among two conferences, the change of an author's research interests, and so on.

EXAMPLE 3.1. *Suppose user specifies two different dimension settings:* ("*author = '*', conference = 'VLDB', year = '1990-1995'* ") *and* ("*author = '*', conference = 'VLDB', year = '1995-2000'* "). *The system then shows the text summaries correspondingly* (*i.e., term frequencies distribution, k topics, and their word distributions*). *The different text summaries from different dimensions essentially show the changes on terms and research topics during the two time periods in "VLDB".*

EXAMPLE 3.2. *Suppose user specifies author = "*", and sets conference = 'ICDE' vs. 'ICDM' to see the difference of topics between these two conferences.*

EXAMPLE 3.3. *Suppose user sets the author dimension to a specific person, and sets conference = "*". By changing the time dimension value, the output of text summaries reflect how this author's research interests changed.*

Second, queries contain both dimension and text information. In this situation, people may be interested in papers on some topic with constraints on other structured dimensions. For example, users may want to find all "data mining" related papers on "VLDB" conference, or all papers related to

Topic 1	Topic 2	Topic 3
mine	graphical	rank
large	graphics	query
index	language	efficient
search	interface	top
structure	design	answer
pattern	process	stream
fast	processor	integration
match	...	visual
...	...	...

Table 2: K topics for query "graph" on user-specified dimension

"cube" for a given author. The computed IR measures in multidimensional space in TextCube and TopicCube greatly facilitate this search. What's more, by using TopicCube, user may get related results that not exactly contain the search text in the query. For example, when user inputs "cube" as the text information in the query, the results may contain "OLAP" and so on.

EXAMPLE 3.4. *Suppose user inputs a query, providing dimension "conference='database conferences', year='*', author='*' " and keyword "graph". The system has two optional outputs depending on which text measure the user wants to use. First, by using topic model as measure, k topics related to text "graph" within the user-specified dimension will be returned. For each topic, top terms are output to describe the topic, as shown in Table 2. And papers in the confined set are output according to the relevance score to the keyword "graph". Second, by using TF/IDF as measure, relevant papers are output according to similarity measure calculated by the vector space model.*

Third, queries only contain text information. This is the same as traditional information retrieval. User can get the results efficiently by either using the inverted index stored in TextCube or by the topic model calculated in the "all" cell in TopicCube.

3.2.2 OLAP

Given the concept hierarchy of the dimensions, which are computed from information network analysis algorithm such as NetClus, OLAP operations can be applied on multidimensional text databases. Different from traditional cube, the measure of each cube is actually text measure, in the form of TF/IDF, or topic model (*i.e.*, term distribution). There are four related OLAP operations *Roll-up*, *Drill-down*, *Slice* and *Dice* [10]. We now give two examples of OLAP operations, namely drill-down and roll-up using DBLP data set as an illustration case.

EXAMPLE 3.5. (*Drill-down on conference dimension*). *Suppose initially user selects conference = "Database and In-*

formation System", other dimensions are all set to "", the system then selects all satisfied papers, and returns the text measure of the set of papers online. Next, user may drill-down on the dimension of conference, e.g., dill down to "Database Conferences" or drill down further to a specific conference, say "VLDB".*

EXAMPLE 3.6. (*Roll-up on author dimension*). *Suppose initially user selects conference = "SIGMOD", author = "Richard T. Snodgrass", and other dimensions are set to "*". Also suppose we already know from* NetClus *the hierarchical clustering result: at first level, Snodgrass belongs to "Database and Information Systems"; at second level he belongs to "Database"; and at third level he belongs to "Temporal Database". Then the user may roll up by setting author to "Temporal Database Researchers" or roll up further to "Database Researchers" or "*"* (*All*).

3.2.3 Mining

Complex mining functions on multidimensional text databases can be further supported by the system platform. For example, how can we find highly ranked author, conference, or year given a user query? How can we find a group of related people given some topic through news data and detect their evolving trend? Notice that, speaking of the "query" here, it could contain both dimension information and text information. In this demo, we only demonstrate a special type of mining task, which is to extract a subnetwork for a given query and apply graph mining tasks and visualization on the subnetwork.

Given a user query, we utilize TextCube and TopicCube to compute the top-k relevant entities for a query. Then we only include those entities and generate a local subnetwork. This small subnetwork is served as a local view extracted from the entire huge information network and may provide better understandings of the query that users are interested in. Furthermore, given a subnetwork, advanced information network analysis tool can be used to find deep and hidden knowledge. Also, visualization of such networks could help user discover interesting patterns.

EXAMPLE 3.7. (*Mining in the DBLP database*). *Given a query with keywords "graph pattern mining", and limit the searching dimension to ("conference = *", "author = *", "year = *"), the system finds the most relevant papers, and a vertex induced subnetwork comprised of relevant authors, conferences, and years will be built. Rankings of authors, conferences and years relative to topic "graph pattern mining" can be calculated using the ranking method similar to the one shown in* NetClus.

4. DEMONSTRATION

In this demo, we will show iNextCube, a user-friendly web-based, information-network enhanced, multidimensional text cube system, on two datasets: the whole DBLP data set and a set of news data. The system uses MS ASP.NET/IIS and SQLServer.

As introduced in Section 3, our system demo on the DBLP dataset will show three functions: search, OLAP, and query-based subnetwork analysis.

For the news data, news agencies usually provide some multidimensional information for each piece of news, such as time, location, news agency, author, news category (*e.g.*, politics → US politics → presidential election → ...), as well as several other dimensions. We can also use classification and clustering methods to extract some other topic hierarchies from the news text. Moreover, entity extraction will be perform based on some term matching, natural language processing, and dictionary consultation methods to extract important entities (*e.g.*, person name and location) to form structured fields and relationships. In addition, topic modeling and topic hierarchy construction methods will be further developed and applied to extraction of additional hierarchies and clusters from the text data. Based on such multidimensional information and the text data, we will perform the NetClus analysis to construct object-ranked clusters and construct TextCube and TopicCube. The so constructed data cube will be used to support search, OALP and data mining.

Our demo on the news data set will select a few popular U.S. and world political news extracted from recent news. The subnetworks containing interrelated entities will be derived to support typical queries like "2008 U.S. president election" by matching it against the textual contents. Subsequently, centrality analysis, clustering, and topic summary will be performed over the subnetwork to demonstrate influential figures, political opinions, *etc.*.

An online iNextCube system will be made publicly accessible at http://inextcube.cs.uiuc.edu.

5. REFERENCES

[1] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao. Text Cube: Computing IR measures for multidimensional text database analysis. In *ICDM*, 2008.

[2] D. Zhang, C. Zhai, and J. Han. Topic Cube: Topic modeling for OLAP on multidimensional text databases. In *SDM*, 2009.

[3] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: Integrating clustering with ranking for heterogenous information network analysis. In *EDBT*, 2009.

[4] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*, 2009.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999

[6] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania. Efficiently linking text documents with relevant structured information. In *VLDB*, 2006.

[7] A. Inokuchi and K. Takeda. A method for online analytical processing of text data. In *CIKM*, 2007.

[8] J. M. Perez, T. B. Pedersen, R. B. Llavori, and M. J. Aramburu. IR and OLAP in XML document warehouses. In *ECIR*, 2005.

[9] J. P. Callan and M. E. Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97-130, 2001.

[10] V. Harinarayan, A. Rajaraman, and J. D.Ullman. Implementing data cubes efficiently. In *SIGMOD Conference*, 1996.