EventCube: Multi-Dimensional Search and Mining of Structured and Text Data

Fangbo Tao[‡], Kin Hou Lei[‡], Jiawei Han[‡], ChengXiang Zhai[‡],
Xiao Cheng[‡], Marina Danilevsky[‡], Nihit Desai[‡], Bolin Ding[‡], Jing Ge[‡], Heng Ji[◇],
Rucha Kanade[‡], Anne Kao[†], Qi Li[◇], Yanen Li[‡], Cindy Xide Lin[‡], Jialiu liu[‡], Nikunj Oza[⋆],
Ashok Srivastava[⋆], Rod Tjoelker[†], Chi Wang[‡], Duo Zhang[‡], Bo Zhao[‡]
[‡] Computer Science, Univ. of Illinois at Urbana-Champaign [⋆] Aviation Safety Group, NASA
[†] Boeing Research & Technology [◇] Computer Science, City University of New York

ABSTRACT

A large portion of real world data is either text or structured (*e.g.*, relational) data. Moreover, such data objects are often linked together (*e.g.*, structured specification of products linking with the corresponding product descriptions and customer comments). Even for text data such as news data, typed entities can be extracted with entity extraction tools. The EventCube project constructs TextCube and TopicCube from interconnected structured and text data (or from text data via entity extraction and dimension building), and performs multidimensional search and analysis on such datasets, in an informative, powerful, and user-friendly manner. This proposed EventCube demo will show the power of the system not only on the originally designed ASRS (Aviation Safety Report System) data sets, but also on news datasets collected from multiple news agencies, and academic datasets constructed from the DBLP and web data. The system has high potential to be extended in many powerful ways and serve as a general platform for search, OLAP (online analytical processing) and data mining on integrated text and structured data. After the system demo in the conference, the system will be put on the web for public access and evaluation.

Categories and Subject Descriptors

H.2.8 [**Information Systems Applications**]: Database Applications—*Data Mining*

Keywords

Data Cube System, Multidimensional Data

1. INTRODUCTION

We are living in the big data age. A large portion of real world big data is either text or structured data. Moreover, such data objects are often linked together (*e.g.*, structured product information linking with their descriptions and customer evaluation). Typically, structured/relational data has been handled by relational database systems, and such systems also provide some text indexing and search capabilities to assist text data stored in such (extended) relational database systems. However, such kind of systems often suffer from the following limitations.

1. It can hardly support systematic search and analysis of large collections of free text in multi-dimensional way, although text data is ubiquitous in real-world;
2. It usually does not support data cube technologies on text data and multidimensional text mining although it is obvious that text mining and data cube technologies can mutually enhance each other; and
3. There is a lack of a general platform that can support integrated multi-dimensional analysis of structured and text data, on top of which many powerful analysis methods and tools can be developed, experimented and refined, such as viewing such data sets as interconnected information networks and further applying information network analysis technology.

EventCube is a project that provides such a general platform that can easily import any collection of free text and structured data, such as news data, aviation reports or academic papers, extract entities, construct the text-rich data cube and support powerful search and mining functions. For structured data, multidimensional data cube can be constructed easily. For text-intensive data with minimally predefined structured information (e.g., news data), natural language and information extraction tools can be used to extract entities of multiple types such as person, location, organization, time, and event. This framework provides a tremendous opportunity to conduct multi-dimensional analysis on text and structured data in powerful and flexible ways. This great potential motivates our research and development of the EventCube system for multidimensional search and analysis of interconnected structured and text data or on text data via dimension construction using entity extraction and dimension building tools.

The EvenCube project has been funded by NASA, developed in the Department of Computer Science, UIUC, as-

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

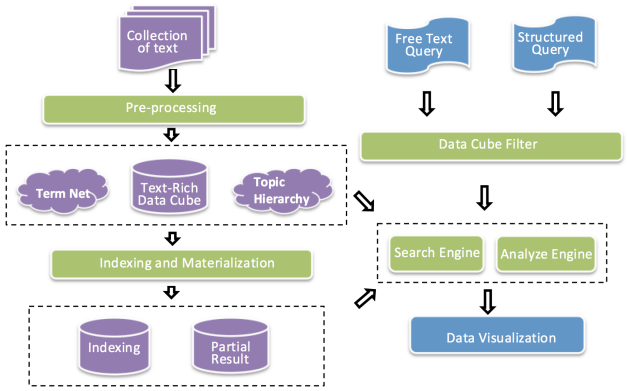

Figure 1: System Architecture of EventCube

sisted by NASA and Boeing researchers, started in 2008, and later it has also been partially supported by ARL (Army Research Lab) via NSCTA (Network Science Collaborative Technology Alliance) program, with many other researchers joined in. With years of research and development, it has reached to a mature stage: Many interesting functions have been developed and integrated into the system and its power can be demonstrated on a spectrum of diverse datasets. The system was originally designed for analysis of ASRS (Aviation Safety Report System) data sets, nevertheless, information extraction tools have been used to extract structured entities from the typical news datasets, collected from multiple news agencies. Therefore, the system becomes a more general platform for search, OLAP (online analytical processing) and data mining on integrated text and structured data.

The system provides multiple search and mining functions with the following architecture.

System Architecture. The EventCube system is designed with the architecture shown in Figure 1. It consists of the following modules: (1) *Data Uploading and Preparation*, which pre-processes the free text corpus from user's uploading and converts it into a text-rich data cube with term network and topic hierarchy extracted; (2) *Indexing and Materialization*, which builds indexing and partial materialization results for keyword search, top cell finding, single dimension distribution and hierarchical topic modeling; (3) *Query-Based Search and Mining Module*, which processes user-queries (both search and analysis queries) by parsing the query, selecting and executing appropriate search or mining module (which searches or mines on the constructed text-rich data cube to derive results); and (4) result presentation by *Visualization and Interpretation* of the search/mining processes and results.

Substantial research have been conducted for the development of the EventCube system, with multiple technologies developed, including TextCube [2], TopicCube [5], TopCells [1], and TEXplorer [6], among others. These technologies have been incorporated into the system. Moreover, multiple powerful, efficient, and flexible search, mining and optimization mechanisms have been incorporated into the system in a user-friendly manner. The system has high potential to be extended in many powerful ways and serve as a general platform for experiments of multidimensional search and mining of text and structured data.

2. MAJOR FUNCTIONAL MODULES

The major functional modules for data preparation, search and mining are described in this section.

2.1 Data Preparation

2.1.1 *Term network construction*

Different from traditional search engines which use exact keyword matching, EventCube generates a term network based on different datasets. This is performed as follows. On any given dataset, the system first extracts frequent terms (including words and short phrases) using typical frequent pattern mining methods. Notice that with frequent pattern mining, the term extraction process can extract phrases with gaps, aggregate those with different orders, and are more efficient and powerful than typical n-gram method based term extraction. Moreover, it automatically maps the abbreviations to the original terms, using expert- or user-provided dictionaries and/or term-mapping datasets. After that, the system captures both the equivalent relationship and correlation relationship at the term level and builds up a term network [4].

2.1.2 *Entity Extraction using NLP and hierarchical topic ontology finding*

From the free text or the text segments of the textual attributes in the integrated structured and text database, NLP tools are used to extract essential entities such as time, location, person, organization and events. Moreover, concept hierarchies (*i.e.*, higher-level entities) are associated with extracted entities (*e.g.*, Chicago is associated with state: Illinois and country: USA) based on user- or expert-provided dictionary or using an entity clustering and concept hierarchy discovery process recently developed in our research [4]. Such a process ensures that the values of the multiple dimensions and the data associated with these dimension values can be automatically or semi-automatically (*e.g.*, with the assistance of user-provided information) extracted, linked, or constructed from free text.

2.2 Search for Text-Rich Data Cube

2.2.1 *Contextual Search*

Compared to traditional search engine, *i.e.*, bag of words model, the search engine in EventCube supports a more intelligent contextual search function. Unlike general search such as Google Search, each search query in our system will only focus on a particular dataset with some particular domain knowledge. Therefore, the system builds a term network which includes the frequent mentioned entities, events and phrases in the corpus. Based on the term-net, we (1) recommend related terms to user's input terms to help improving their queries, (2) support AND, OR, NOT three boolean operators to compose advanced query, and (3) include the equivalent terms, *e.g.*, abbreviations, as a part of query by default.

Besides query for short terms, EventCube also supports bulky text search. Inputting a long text, the system will extract frequent terms from the text, locate them in the term network, improve the query by adding equivalent terms and highly related terms. Based on this principle, the 'Similar Docs Search' function in EventCube can have the ability

Rank	Year	Event	Organization	#Document	Avg-Rele
1	201001	Movement:Transport	White House	8	12.995336890220642
2	201001	Life:Die	White House	6	11.930548350016275
3	201001	Movement:Transport	Navy	4	13.26450800895691
4	201001	Personnel:Elect	Congress	3	7.3021542231241865
5	201001	Contact:Meet	White House	3	13.659941037495932
6	201001	Personnel:Start-Position	United Nations	3	13.533504803975424

Figure 2: Top Cell List for keyword 'Haiti Earthquake' on dimensions of 'Year', 'Event', 'Org' in the News data

to find semantically similar docs, even when they contain totally different words distribution.

2.2.2 *Top-Cell Finding*

A cell in the text cube aggregates a set of documents with matching dimension values on a subset of dimensions. Given a keyword query, our goal is to find the top-k most relevant cells in the text cube. A relevance scoring model and efficient ranking algorithm has been proposed in [1]. It optimizes the search order and prunes the search space by estimating the upper bounds of relevance scores in the corresponding subspaces, so as to explore as few cells as possible for finding top-k answers. An example on the News dataset to generate top-k cells is shown in Figure 2.

2.2.3 *Single Dimension Distributions based on Keywords*

For each search query, it is desirable to provide many insights for analyzers if the data distribution can be provided on each dimension. In EventCube, we aggregate the relevant documents on every dimension to show the heatmap of the keywords for geographic dimension, time series of the keywords for time dimension and the ranked list for other dimensions. An example on aviation safety data can be found in Figure 3.

To achieve both efficiency and effectiveness, we propose a framework that combines offline and online computation together to generate real-time single dimension distribution results for every query.

In the offline part, we first map equivalent terms into one, then build both keyword-doc inverted index and cell-doc inverted index, finally merge them to calculate single term distribution for each pair of terms and dimensions. In the online part, we match each term in user's query into its equivalent term. Based on the assumption of independency of terms, we use De Morgan's laws to estimate the combined term-dimension distribution.

2.3 Analysis for Text-Rich Data Cube

2.3.1 *Hierarchical TopicCube Generation*

Probabilistic topic models are among the most effective approaches to latent topic analysis and mining on text data. On the other hand, online analytical processing (OLAP) techniques are useful for analyzing and mining structured data, such as data cubes. By combining OLAP and topic modeling together, we treat hierarchical topic distribution as one of the aggregation functions. With the support of

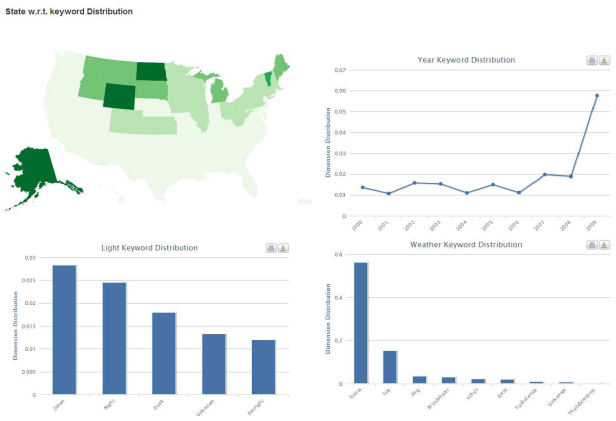

Figure 3: Single Dimension Distribution for keyword 'Snow' on the Aviation Safety Reporting Data

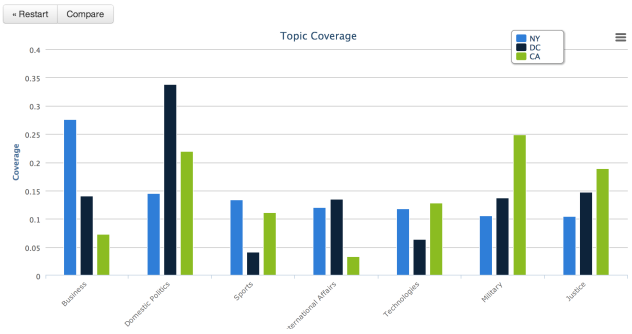

Figure 4: Topic Distribution Comparison for Cells "NY", "DC", and "CA"

comparison of topic distribution for different cells, our system can provide deeper insights for decision makers. This is realized efficiently by our TopicCube model [5] that constructs a hierarchical topic tree on data cube to define a topic dimension for exploring text information. An example on news data is shown in Figure 4.

To improve the TopicCube model, EventCube also generates n-grams to describe the topic. The underneath method is proposed in [3].

3. ABOUT THE DEMO

The EventCube system provides an easy-to-use web interface and easy-to-understand, elegant data visualization. Figure 5 and 6 is a preliminary screen shot of an example search result.

In this demo, we build a complete automatic workflow to analyze a collection of text data. A user can easily browse different datasets to be analyzed in a dataset portfolio page. Also, she can upload their particular text corpus in an easy-to-use way. The system then creates an independent thread to extract entities/dimensions, construct the term-net, find the topic hierarchy, build indexes and compute partial materialization results as a background process. When the preprocessing is done, the system will inform the user and make the dataset accessible. A user is then permitted to conduct search and analysis tasks.

This demonstration will show three sample datasets: (1) NASA's Aviation Safety Reporting System data, which contains 60,000 US. aviation reports from pilots or maintenance

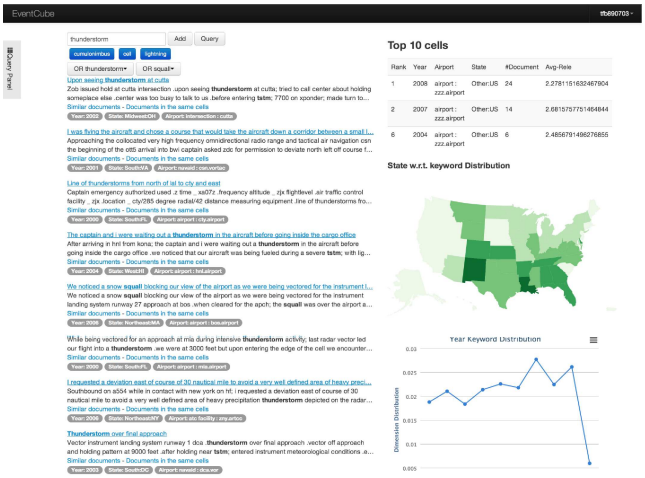

Figure 5: Snapshots for Search Result Page with Keyword 'thunderstorm'

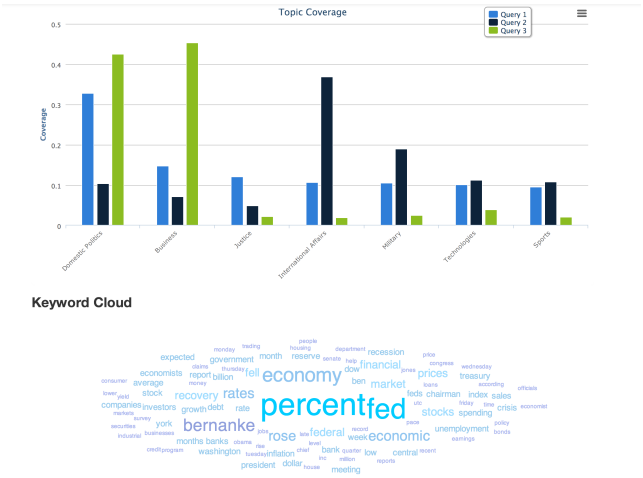

Figure 6: Snapshot for Analysis Result Page. A comparison between "Barake Obama", "Hillary Clinton" and "Ben Bernanke"

workers. For each report, it has 11 dimensions including Year, Airport, State, etc.; (2) News Data, which contains 500,000 US domestic news articles from 2007 to 2011. For each news article, it has 6 dimensions including Time, Location, Agency, Event, People and Organization; and (3) Academic Data, which contains 100,000 computer science publications from 1975 to 2012. However, since the components and functionalities of EventCube are not confined to these demonstrated datasets, it will be convenient for other researchers to try their own datasets with the system and extend the capability of EventCube. Since different text data may refer to different domains, which may require different OLAP and data mining functionalities as well different entity extraction and topic ontology finding functions, much research is needed to extend such an EventCube system to become a general, powerful and intelligent text-based search and mining platform.

A strong motivation for this study is to show the power of analyzing massive free text data by combining data mining, data cube and natural language processing technologies. Mining such data will generate useful knowledge and provide deep insights for decision makers. We hope this demo may serve this purpose and promote the development of more powerful mining methods to tackle such big text data.

For interested readers who would like to play with the system, please check www.eventcube.org, with kdd2013/kdd2013 as username/password to login. Please note the system is still under active evolution with new functions incrementally added in and/or revised. Thus the description outlined here may not be exactly matching the functionality of the system (which will be richer), depending on the time when you try it online.

4. REFERENCES

[1] B. Ding, B. Zhao, C. X. Lin, J. Han, C. Zhai, A. Srivastava, and N. C. Oza. Efficient keyword-based search for top-k cells in text cube. *IEEE Trans. on Knowledge and Data Engineering (TKDE)*, 23:1795–1810, 2011.

[2] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao. Text Cube: Computing IR measures for multidimensional text database analysis. In *Proc. 2008 Int. Conf. Data Mining (ICDM'08)*, Pisa, Italy, Dec. 2008.

[3] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 649–655, Philadelphia, PA, Aug. 2006.

[4] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. A phrase mining framework for recursive construction of a topical hierarchy. In *Proc. 2013 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'13)*, Chicago, IL, Aug. 2013.

[5] D. Zhang, C. Zhai, J. Han, A. Srivastava, and N. Oza. Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications. *Statistical Analysis and Data Mining*, 2:378–395, 2009.

[6] B. Zhao, J. Han, C. X. Lin, and B. Ding. TEXplorer: Keyword based object ranking and exploration in multidimensional text databases. In *Proc. 2011 Int. Conf. on Information and Knowledge Management (CIKM'11)*, Glasgow, UK, Oct. 2011.

Acknowledgments. Work supported in part by NASA under grant NNX08AC35A, Boeing Corp., the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. National Science Foundation grants IIS-0905215, CNS-0931975, and IIS-1017362, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, DTRA, and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the of?cial policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.