NewsNetExplorer: Automatic Construction and Exploration of News Information Networks

Fangbo Tao[†], George Brova[†], Jiawei Han[†], Heng Ji[‡], Chi Wang[†], Brandon Norick[†], Ahmed El-Kishky[†], Jialu Liu[†], Xiang Ren[†], Yizhou Sun[◇]

[†] Department of Computer Science, University of Illinois at Urbana-Champaign
[‡] Department of Computer Science, Rensselaer Polytechnic Institute
[◇] College of Computer and Information Science, Northeastern University

ABSTRACT

News data is one of the most abundant and familiar data sources. News data can be systematically utilized and explored by database, data mining, NLP and information retrieval researchers to demonstrate to the general public the power of advanced information technology. In our view, news data contains rich, inter-related and multi-typed data objects, forming one or a set of gigantic, interconnected, heterogeneous information networks. Much knowledge can be derived and explored with such an information network if we systematically develop effective and scalable data-intensive information network analysis technologies.

By further developing a set of information extraction, information network construction, and information network mining methods, we extract types, topical hierarchies and other semantic structures from news data, construct a semi-structured news information network NewsNet. Further, we develop a set of news information network exploration and mining mechanisms that explore news in multi-dimensional space, which include (i) OLAP-based operations on the hierarchical dimensional and topical structures and rich-text, such as cell summary, single dimension analysis, and promotion analysis, (ii) a set of network-based operations, such as similarity search and ranking-based clustering, and (iii) a set of hybrid operations or network-OLAP operations, such as entity ranking at different granularity levels. These form the basis of our proposed NewsNetExplorer system. Although some of these functions have been studied in recent research, effective and scalable realization of such functions in large networks still poses multiple challenging research problems. Moreover, some functions are our on-going research tasks. By integrating these functions, NewsNetExplorer not only provides with us insightful recommendations in NewsNet exploration system but also helps us gain insight on how to perform effective information extraction, integration and mining in large unstructured datasets.

SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.

http://dx.doi.org/10.1145/2588555.2594537.

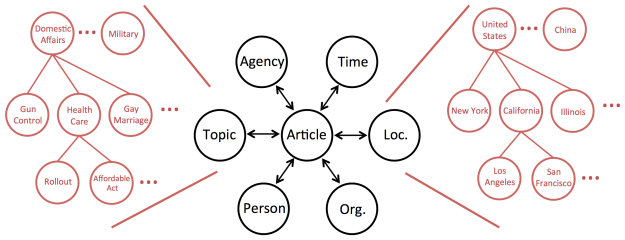

Figure 1: News Network Schema: The hierarchies of "*Topic*" and "*Location*" are illustrated in the graph, the hierarchies of "*Person*", "*Time*" and "*Organization*" are not shown for simplicity.

Categories and Subject Descriptors

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

Keywords

Information Network Construction, Network-OLAP

1. INTRODUCTION

We are living in the dawn of big data era. Massive amount of data has been generated in fast pace from every corner of our society. It is important for database and data mining researchers to demonstrate the promise of our technology using massive real datasets. News data is one of the most abundant and well-understood data sources. It is ideal to explore news data thoroughly and use it to systematically demonstrate the power of information technology. Unfortunately, unlike research publication data sets (*e.g.*, DBLP) or most relational database data, news data is largely unstructured. It is a big challenge to turn such unstructured news data to semi-structured data automatically.

In our recent research, we have made progress in two frontiers. First, we have systematically studied methods for data mining in heterogeneous information networks [4] and developed a set of methods for mining heterogeneous information networks. Taking computer science bibliographic networks extracted from DBLP as an example, one can (i) cluster and rank venues, authors and terms in computer science by RankClus and NetClus [4], (ii) derive quality classification models for multi-typed entities and present their ranking as by-product by GNetMine and RankClass [3], (iii) conduct effective similarity search across networks by PathSim [5], and

*etc.*Clearly, modeling heterogeneous datasets as structured heterogeneous information networks often captures richer node/link semantics and generates better mining and search results than their underlying relational databases. Such mechanisms can be transferred to news data analysis if the multi-dimensional structured information can be extracted from news data and a heterogeneous, semi-structured news information network can be constructed automatically.

Second, we have been developing effective methods for constructing heterogeneous information networks from unstructured text data. We have studied how to extract typed data from news data with advanced NLP technology, which forms the base for construction of a preliminary heterogeneous information network. Moreover, we have recently developed a phrase mining framework for recursive construction of topical hierarchies from text data [10] and from heterogeneous information networks [11]. By further refinement of the above processes, a quality NewsNet will be able to be generated systematically.

In this demo, we will show how a quality NewsNet can be constructed by integration of NLP, phrase mining, and information network analysis methods, and how OLAP and heterogenous network mining can be successfully performed on NewsNet. NewsNetExplorer is our new project based on our recent research and based on experience on the development of the two previous demo systems: Research-Insight on the DBLP data [9] and EvenCube on NASA's Aviation Safety Report data [8].

2. NETWORK CONSTRUCTION

The news corpora released through Linguistic Data Consortium (**LDC**) to research communities is a rich news collection to build a comprehensive news network. It includes 10 million news articles from 7 different major news agencies in the Gigaword Corpus[1]. For such an enormous collection of plain text articles, the very first process we apply is extracting typed entities from text and construct an interconnected network with sophisticated hierarchies associated to these typed entities. As shown in Figure 1, we build a network, generated by taking *Article* as a center type, linking to multiple typed nodes including *Person, Organization, Agency, Time, Location* and *Topic*. Moreover, except the center type *Article* and *Agency*, other typed nodes are organized in a hierarchical way as illustrated in Figure 1.

2.1 Entity extraction using NLP

From the free text or the text segments of the textual attributes in the integrated structured and text news data, natural language processing (**NLP**) (especially Information Extraction) tools are used to extract essential entities such as time, location, person, organization. Moreover, concept hierarchies (*i.e.*, higher-level entities) are associated with extracted entities (*e.g.*, Chicago is associated with state: Illinois) based on a user- or expert-provided dictionary.

2.2 Entity extraction using hierarchical topic ontology finding

For *Topic* extraction of NewsNetExplorer we have developed a method[10] that construct a topical hierarchy from a collection of text. The framework called **CATHY** (Construct A Topical HierarchY) is a recursive clustering and ranking approach for topical hierarchy generation. In the news collection, many pieces can describe the same topic; meanwhile, different topics may reflect the same fact at different levels of granularity. The aim of [10] is to construct a hierarchy where each topic is represented by a ranked list of phrases, such that a child topic is a subset of its parent topic. This strategy also works well specifically for news data since news articles tend to use different phrases to report the same topics.

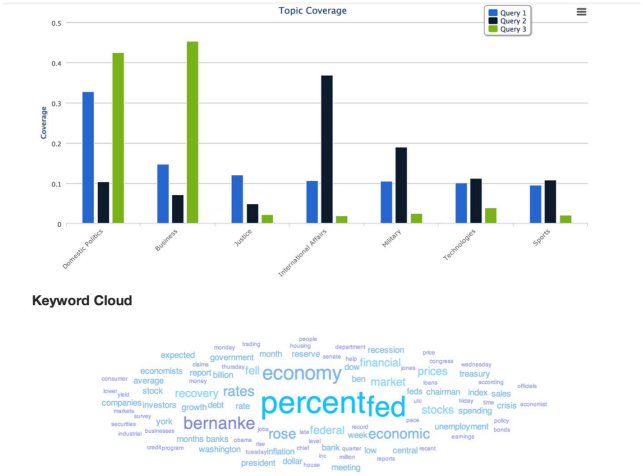

Figure 2: A comparison of topic distributions between "Barack Obama", "Hillary Clinton" and "Ben Bernanke". The tag cloud is a summary for Ben Bernanke's news articles.

3. MAJOR FUNCTIONAL MODULES

By extracting the entities from millions of news articles and building highly sophisticated hierarchies for those typed entities, we can view the structured information from two different perspectives: (1) Heterogeneous Information Network with Rich Text, and (2) Multi-dimensional Hierarchical Data Cube with Rich Text. Clearly, modeling a heterogeneous network as Figure 1 captures richer node/link semantics and generate better mining result than a homogeneous network or unstructured text collection. Meanwhile, a multi-dimensional data cube enables us to carry out advanced online analytical processing (**OLAP**) operations at different levels of granularity. We also explore a new direction of leveraging both information network and multi-dimensional data cube structures, conducting network-based mining algorithms at different levels of granularity of the network. The major functional modules are described in this section.

3.1 OLAP Operations on News Data

3.1.1 Hierarchical Cell Summary

OLAP techniques are efficient and effective for mining and analyzing structured data like data cubes. With rich text gained from millions of news articles, it also becomes possible to mine the text data and analyze latent topics in an OLAP fashion. By combining OLAP and probabilistic topic modeling together, we treat topic distribution of a set of news articles as an aggregation function. By exploring at different granularity levels and comparing topic distributions for different cells, we can reveal deeper insights in the huge news collection. The algorithm implemented here is proposed by our TopicCube model [13] that constructs

Rank	Cell	#Document	Avg-Relevance
1	Time:Nov-2009, Organization:House of Rep.	93	2.1201770279997136
2	Time:2010, Organization:White House	51	1.9302240668558608
3	Time:2013, Topic:Healthcare.gov rollout	20	1.9351981639862061

Figure 3: Top-ranked cells for the query "Healthcare", the top results indicate 1). House of Representatives passed the bill, 2). The President signed the bill. 3). The website rollout.

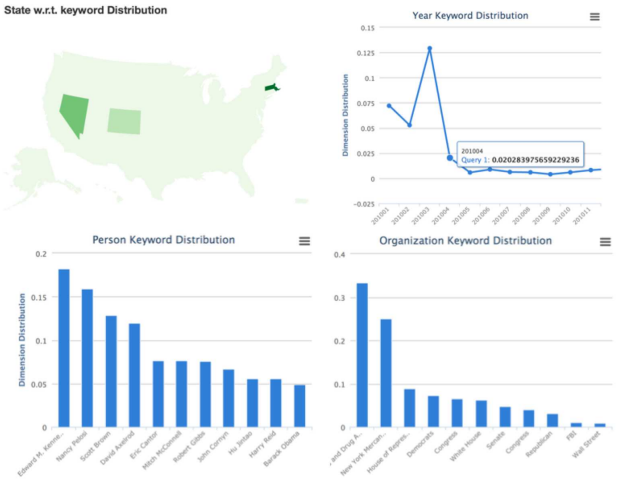

Figure 4: Single dimension distribution for query "healthcare bill". We find top related *Location, Time, Person and Organization.*

a hierarchical topic tree on a data cube to define a topic dimension for exploring text information. An example on news data is shown in Figure 2.

3.1.2 Promotion Analysis in News Cube

A cell in the text cube aggregates a set of documents with matching dimension values on a subset of dimensions. Given a keyword query on the news, we want to enable people to find most relevant cells in the data cube. A most relevant cell needs to be generated from two aspects: (1). It is a combination of relevant dimensions. (2). It needs to find the best level of granularity of each dimension to describe the cell. For a query like "Healthcare bill", we may need to find a cell as "{Organization: Congress, Time:2009}", while for a query like "Nelson Mandela", a "{Time:Dec/2013, Topic:Death}" cell is better. The top-ranked cells should not only be highly relevant, but also be significant for the query. A relevance scoring model and efficient ranking algorithm has been proposed in [2] and [12]. 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 3.

3.1.3 Hierarchical Single Dimension Distributions

For each news search query, it is desirable to provide many insights for analysts if the data distribution can be provided on each dimension. For instance, if people want to know more about the *healthcare bill*, other than providing a list of relevant documents for them to read, it's helpful to show top relevant entities on *Person, Organization, Time* and *Topic* dimensions. This structured summary can also be hierarchical, which will enable users to explore along different levels of the *Time, Topic* or *Location* hierarchies. In NewsNetExplorer we aggregate the relevant documents on every dimension to show the heatmap of the keywords for *Location* dimension, time series of the keywords for *Time* dimension and the ranked list for other dimensions. An example on query "Healthcare Bill" can be found in Figure 4. 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.

3.2 Information Network Mining Operations

3.2.1 Similarity Search

Similarity search often plays an important role in the analysis of networks. For a news collection we use, which contains millions of articles covering several decades of various news stories, it's critical and challenging to find the connections between entities in the network. By considering different linkage paths in a network, one can derive various semantics on similarity. A meta-path based similarity measure is introduced [5], where a meta-path is a structural path defined at the meta level (*i.e.*, relationships among object types). It turns out to be more meaningful in many scenarios compared with random-walk based similarity measures and is also efficient for top-k similarity search in heterogeneous networks. In NewsNetExplorer we implemented this similarity algorithm in a brand new scenario of news network.

Example 1: Similarity Search in NewsNetExplorer Given a person (*e.g.*, Barack Obama), find his/her top-k similar people and explain why (by summarizing their connections and the corresponding similarity measure). We expect to find other presidents by meta-path *P(erson) - O(organization) - P(erson)* and to find most relevant contemporary politicians by meta-path *P(erson) - T(opic) - P(erson)*. Do the same for an organization (*e.g.*, Senate), a state (*e.g.*, Illinois) and a topic(*e.g.*, Iran nuclear crisis). Potential extensions include finding top-k most related heterogeneous typed objects (*e.g.*, given a person (*e.g.*, Lady Gaga), find his/her top-k most related organizations and topics). ■

3.2.2 Ranking-based Clustering

Suffering from the complexity of news articles, it has always been important to apply clustering based on news content. More than traditional approaches, we apply a link-based clustering algorithm to utilize the structured network, which explores links across heterogeneous types of data. Our recent studies develop a ranking-based clustering approach, represented by RankClus [6] and NetClus [7], that generates interesting results for both clustering and ranking efficiently. A significant difference from traditional text clustering is that we also cluster typed entities like *Person, Topic and Organization*. This approach is based on the observa-

tion that ranking and clustering can mutually enhance each other because objects highly ranked in each cluster may contribute more towards unambiguous clustering, and objects more dedicated to a cluster will be more likely to be highly ranked in the same cluster.

Example 2: Rank-based clustering in NewsNetExplorer· Given a subnetwork (*e.g.*, network formed under specific conditions, *e.g.*, years 2005-2008) and a desired number of clusters (*e.g.*, 4), perform rank-based clustering and show top-k objects of each type (*i.e.*, *Person, Topic, Organization*) in each cluster. ■

3.3 Network-OLAP Operations

3.3.1 Entity Ranking on Different Granularity Levels

Entity Ranking is an important feature to help readers understand different news lines better without going through the related articles. People are interested in questions like "*who are the most important people in 2013*", "*what are the most relevant organizations about 911 attack?*" and "*what's the major topics of China during 2000 to 2010?*". To be expressed in our framework, given a cell of the data cube, find the top-ranked entities under the cell condition. We applied a linkage-based ranking algorithm here, in which we first build the subnetwork according to the cell, then apply the ranking function for different kinds of entities in the sub-network. We introduced a two-step approach to combine both global ranking and local ranking together into the ranking function. We treat a graph algorithm as our aggregation function, that traditional OLAP algorithms cannot be applied to improve the efficiency.

Example 3: Entity Ranking in NewsNetExplorer· Given a cell (*e.g.*, *2013 and Iran Nuclear Crisis*), perform and show top-k objects of each type (*i.e.*, *Person, Topic, Organization, Time and Location*). Drill down to {*2013-Nov, Iran Nuclear Crisis*} or roll up to {*2013, International Affairs*} to see the new ranking result for each type. ■

4. ABOUT THE DEMO

In this demo, we build a complete automatic workflow to analyze a collection of news articles. We build a text-rich news information network by extracting the typed entities and their hierarchical structure from the *Gigaword* news collection [1], implement three different kinds of mining or search operations based on myriad research work. Users can interactively make their structured queries or text queries to explore the news world in an efficient and effective way.

5. REFERENCES

[1] `http://catalog.ldc.upenn.edu/LDC2011T07`.

[2] 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.

[3] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proc. 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, CA, Aug. 2011.

[4] Y. Sun and J. Han. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.

[5] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proc. 2011 Int. Conf. Very Large Data Bases (VLDB'11)*, Seattle, WA, Aug. 2011.

[6] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In *Proc. 2009 Int. Conf. Extending Data Base Technology (EDBT'09)*, Saint-Petersburg, Russia, Mar. 2009.

[7] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proc. 2009 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, June 2009.

[8] F. Tao, K. H. Lei, J. Han, C. Zhai, X. Cheng, M. Danilevsky, N. Desai, B. Ding, J. Ge, H. Ji, R. Kanade, A. Kao, Q. Li, Y. Li, C. X. Lin, J. liu, N. Oza, A. Srivastava, R. Tjoelker, C. Wang, D. Zhang, and B. Zhao. Eventcube: Multi-dimensional search and mining of structured and text data. In *Proc. 2013 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'13)*, Chicago, IL, Aug. 2013.

[9] F. Tao, X. Yu, K. H. Lei, G. Brova, X. Cheng, J. Han, R. Kanade, Y. Sun, C. Wang, L. Wang, and T. Weninger. Research-insight: Providing insight on research by publication network analysis. In *Proc. of 2013 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'13)*, New York, NY, June 2013.

[10] 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.

[11] C. Wang, X. Yu, Y. Li, C. Zhai, and J. Han. Content coverage maximization on word networks for hierarchical topic summarization. In *Proc. of 2013 Int. Conf. on Information and Knowledge Management (CIKM'13)*, San Francisco, CA, Oct. 2013.

[12] T. Wu, Y. Sun, C. Li, and J. Han. Region-based online promotion analysis. In *Proc. 2010 Int. Conf. on Extending Data Base Technology (EDBT'10)*, Lausanne, Switzerland, March 2010.

[13] 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.

Acknowledgments. This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, U.S. NSF CAREER Award under Grant IIS-0953149, U.S.DARPA Award No. FA8750-13-2-0041 in the Deep Exploration and Filtering of Text (DEFT) Program, U.S. National Science Foundation grants CNS-0931975, IIS-1017362, IIS-1320617, IIS-1354329, DTRA, MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC, a Microsoft Research Gift, and an NSF Graduate Fellowship, IBM Faculty award and RPI faculty start-up grant.