A Flexible Topic-driven Framework for News Exploration

Juanzi Li, Jun Li, and Jie Tang
Department of Computer Science and Technology
Tsinghua University, 100084. China
{ljz, lijun, tangjie}@keg.cs.tsinghau.edu.cn

ABSTRACT

With the flourishing of various Web applications, the Internet has become one of the most important means to access news. According to one investigation, in the population of Internet users, 78.5% are looking for news. Unfortunately, although the Internet provides a platform for easily sharing information, it also brings a fast explosion of the news data. It leads to the fact that people spend more and more time to digest the data. Can we design new ways to help the users quickly understand and explore the news data? Information retrieval is one way, but it is insufficient.

In this paper, we propose a flexible topic-driven framework, namely NewsInsight, for news exploration. This framework innovatively integrates a probabilistic topic model, graphical data analysis, and natural language processing. It performs news mining at the topic level and presents news information with topics, entities (e.g., people, organization, and events), and relations derived from the news data. Based on this framework, we have developed a system which can help people to understand and explore news from multiple dimensions. The trial operation of the system has worked at Xinhua News Agency, one of the biggest news publishers in China. Feedback from users shows that the system has achieved its primary objectives.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining; H.2.8 [**Database Management**]: Database Applications

General Terms

Algorithms, Experimentation

Keywords

News Mining, Topic Model, Visualization, Data Exploration

1. INTRODUCTION

With the emergence and rapid proliferation of Web applications, in particular social Web applications, such as content providing sites (e.g., Yahoo!, CNN.com, WallStreet), social sharing sites (Digg, Wikipedia, Facebook, MySpace), and blogs (e.g., Blogger, WordPress, LiveJournal), and microblogs (e.g., Twitter, Jaiku), to mention a few, there is little doubt that we are facing an era of data explosion. News, as one of the most common means for sharing information, has already become necessary for most people in their daily life. A statistic from CNNIC shows that about 234 millions users in China(78.5% of all Chinese Internet users) rely on the Internet to acquire news information. Statistics also unveil that people are spending more and more time to digest news data, due to the fast explosion of the volume of the data. Therefore, there is a clear need for methods and techniques to help users to easily understand and explore news data.

Prior work on news mining/management can be mainly categorized into two classes: document based and word based [6]. The former takes each news item as a unit and aims to organize the news into a well-defined structure. For example, classification and clustering techniques have been used to classify the news into predefined or automatically discovered categories. The latter work, i.e., word based, mainly focuses on the word level, for example discovering the high frequency terms or identifying the hot query keywords. Unfortunately, these methods are insufficient for exploring news. Recently, Microsoft has developed a system, called Renlifang[1], to mine the relationship between people extracted from the news. However, it still does not consider in-depth analysis of the news insights, for example, how to extract the 5W1H (When, What, Who, Where, Why, and How) [5] information from news, how to automatically discover the topics presented in the news, how to design a schema to formally describe the extracted information, and how to visualize the information of multiple different views discovered from news.

We argue that such an in-depth analysis is the core function that is missing in many existing news mining systems. 5W1H are the basic elements of news. Topic discovery can uncover what a news item talks about, and identify which news documents talk about the same topic. With this discovered information, a news reader can first have a view of what topics are presented in today's news. He can zoom in on a preferred topic and take a look at the major elements (5W1H) in the topic. Essentially, he first gets summarized information on the topic. If he is really interested in the topic, he can finally take a close look at the news most related to the topic. Similarly, the in-depth analysis can also benefit many other applications such as news event detection.

To achieve the above goals, we propose a flexible topic-driven framework for news exploration, namely NewsInsight. The key features of NewsInsight are the following:

KDD'09, June 28–July 1, 2009, Paris, France.
ACM 978-1-60558-495-9/09/06.

[1] http://renlifang.msra.cn

- NewsInsight presents a unified framework for news exploration, while previous news mining work usually deals with the different issues in a separated way;
- Topics, entities and relations are proposed to represent the news in a statistical and semantic way, which can support multi-view exploration for news;
- A visualization engine provides an intuitively and lively graphical user interface to access news.

2. NEWSINSIGHT

NewsInsight is a topic-driven analysis and visualization tool that offers both overview and in-depth exploration techniques for news data.

2.1 Formulation

The goal of NewsInsight is to provide a comprehensive analysis of news data. First we introduce some terminology and definitions, then define the topic-driven news analysis problem.

Topic model: Generally, a news document usually consists of multiple topics. For example, a financial news document may be about topics including “American economy collapsing”, “How the US government deals with the problem”, and “Effect of the collapse on the US financial system”. Formally, for each news $d \in D$, we associate a vector $\theta_d \in \mathbb{R}^T$ of T-dimensional topic distribution ($\sum_t \theta_{dt} = 1$). Each element θ_{dt} is the probability of the news document describing the topic t.

Entity: Entities are the key components to identify the 5W1H in news. We define four kinds of entity: Time, Person, Location and Organization. Person and Organization represent “Who” and “What”, Location "Where", and Time "When", respectively. Normally, there are multiple entities in one news document. We define the entities which satisfy $p(e_i|z) \geq \epsilon$ as relevant entities to the topic z.

Relation: Topics and entities are not independent of each other in news documents. Relations between them can help people understand what topics and entities are related and how they are related. We consider three kinds of relations: topic correlation, entity relation and the relation between topics and entities. They are defined as $R = \{(\phi, L, o_i, o_j)\}$, where o_i and o_j can be the topic or entity, and ϕ and L are the weight and label of the relation.

The formulation provides a way to explore news data from multiple different views such as the major topics, the hot entities and their relations discussed in the news. Benefiting from the multiple views, the user can either quickly understand the main idea or gain a detailed insight of the news data.

Based on the above concepts, we define the tasks of topic-driven news exploration as follows: Given a set of news documents $D = \{d_i\}_{i=1}^{|D|}$, how to extract the topic distribution for each news document? How to identify the relations between entities (or topics)? How to discover the change trend of topics and the entities? Further, how to visualize the discovered topic-level analysis results?

2.2 Architecture

Figure 1 shows the architecture of the NewsInsight system. The system consists of four main components.

1. *Preprocessing*: it includes two main sub-components: tokenization and stop word filtering. For English text, tokenization is done by separating the text using symbols like space, line break, and period. For Chinese text, tokenization would be a bit difficult. We use a tool called ICTCLAS[4][2] to tokenize the Chinese text. After tokenization, we also identify the named entities such as Time, Person, Organization, and Location. Stop word filtering is to remove the high frequency words such as “a” and “the”. For the Chinese text, it removes empty words, modal words, and some adverbs.

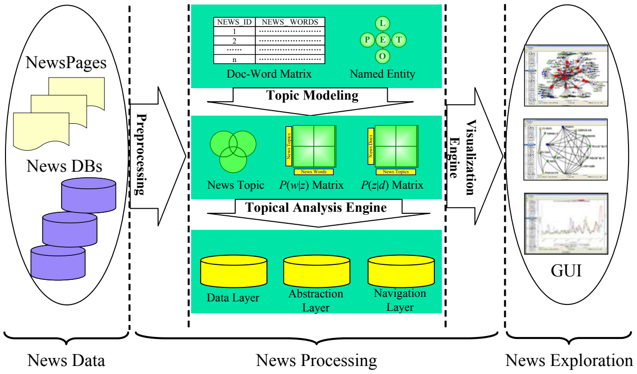


Figure 1: Architecture of NewInsight.

2. *Topic modeling*: this is the basic component in the system. Theoretically, the topic information can be obtained in many different ways. For example, one can use the predefined categories for each news item as the topic information. In addition, we can use statistical topic models [3][1] to automatically extract topics from the news data.

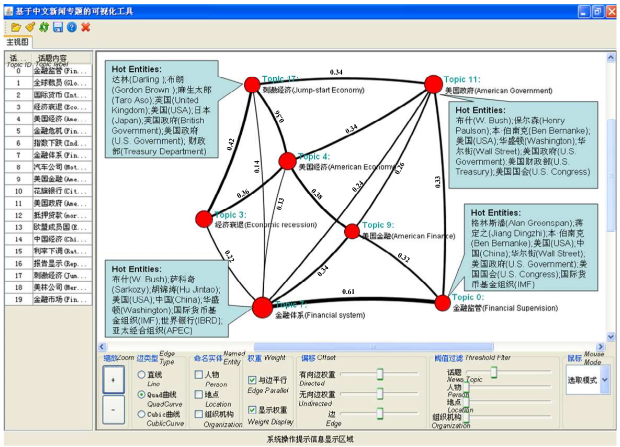
3. *Topical analysis engine*: based on the extracted topic model, the analysis engine performs the comprehensive understanding for news data from different granularities. It has four main functions. It can derive the analysis for topics and entities respectively; it calculates the relations between topics, entities, the relations between topics and entities; and it can model the change trends of different topics and entities. After analysis, we can organize the news documents from three different levels of data(documents and words), abstraction(topics, entities and relations) and navigation(connecting news documents with topics, entities and relations).

4. *Visualization engine*: this presents the discovered topic model and the topical analysis results to the end users in a visualized view. Basically, it includes three types of views: overview view, entity and relation view, topic and entity trend view. The overview view provides a topic based overview information of the news data. The entity and relation view shows the named entities extracted from the news and the relations between them, while the topic and entity trend view gives their change trend.

2.3 Topic Modeling

We employ LDA [1] to model the topics of news documents. Intuitively, in the topic model, each news document is generated by following a stochastic process: first the author would decide what topic z to write according to $p(z|d)$, which is the topic distribution of the document. Then a word w_{di} is selected to represent the semantics of the topic z according to the word distribution of the topic $p(w|z)$. Thus, for training the model, the task is to estimate the parameters: (1) the distribution θ of $|D|$ document-topics and the distribution ϕ of $|T|$ topic-words; (2) the corresponding topic z_{di} for each word w_{di} in the document d. We use Gibbs sampling for

[2] http://www.nlp.org.cn/project/project.php?proj_id=6



话... Topic ID	话题内容 Topic label
0	金融监管 (Fin...
1	全球裁员 (Glo...
2	国际货币 (Int...
3	经济衰退 (Eco...
4	美国经济 (Ame...
5	金融危机 (Fin...
6	指数下跌 (Ind...
7	金融体系 (Fin...
8	汽车公司 (Mot...
9	美国金融 (Ame...
10	花旗银行 (Cit...
11	美国政府 (Ame...
12	抵押贷款 (mor...
13	欧盟成员国 (E...
14	中国经济 (Chi...
15	利率下调 (Rat...
16	报告显示 (Rep...
17	刺激经济 (Jun...
18	美林公司 (Mer...
19	金融市场 (Fin...

Hot Entities:
达林(Darling);布朗(Gordon Brown);麻生太郎(Taro Aso);英国(United Kingdom);美国(USA);日本(Japan);英国政府(British Government);美国政府(U.S. Government); 财政部(Treasury Department)

Hot Entities:
布什(W. Bush);保尔森(Henry Paulson);本·伯南克(Ben Bernanke);美国(USA);华盛顿(Washington);华尔街(Wall Street);美国政府(U.S. Government);美国财政部(U.S. Treasury);美国国会(U.S. Congress)

Hot Entities:
格林斯潘(Alan Greenspan);蒋定之(Jiang Dingzhi);本·伯南克(Ben Bernanke);美国(USA);中国(China);华尔街(Wall Street);美国政府(U.S. Government);美国国会(U.S. Congress);国际货币基金组织(IMF)

Hot Entities:
布什(W. Bush);萨科奇(Sarkozy);胡锦涛(Hu Jintao);美国(USA);中国(China);华盛顿(Washington);国际货币基金组织(IMF);世界银行(IBRD);亚太经合组织(APEC)

Figure 2: News topics and Correlations.

parameter estimation. For details, please refer to [2]. After Gibbs sampling, we can use the sampled topics for words to estimate the probability of a topic given a document θ_{dz} and the probability of a word given a topic ϕ_{zv}.

2.4 Topical Analysis

In topical analysis, we develop three components and implement a corresponding tool to visualize the analysis results. To better illustrate our work, we use the financial news about "economic crisis" from Sina.com[3] as an example in this paper.

News Topics and Correlations

After topic modeling, we get $p(z|d)$ and $p(w|z)$. Now, a common, major challenge in applying such topic models to any text mining problem is to label a topic model accurately so that a user can interpret the discovered topics. So far, such labels have been generated manually in a subjective way. However, it is not only time-consuming but also inaccurate.

In NewsInsight, for labeling a topic, we first select 20 words with high generative probabilities $p(w|z)$ for the topic as the candidates S_1. Then, based on S_1, we generate the set S_2 of word bigrams by combining words in S_1 according to the predefined patterns. Typically, in topic labeling, we mainly consider noun phrases and verb phrases to be the candidate labels. Next, we calculate a score for each candidate label in S_2 by:

$$Score^i(\vec{w}) = \sum_{w \in \vec{w}} p(w|z_i) \cdot tf^i(w) \cdot idf^{-i}(w) \qquad (1)$$

where, $tf^i(w)$ denotes the frequency of candidate w occurring in all documents relevant to topic z_i, and $idf^{-i}(w)$ is the inverse of frequency of w occurring in all documents of other topics except topic z_i. Finally, we select the label with the maximal $Score^i(\vec{w})$ as the label for topic z_i.

For approximating the correlation between topics, we use a cosine-similarity based method to define the correlation between topics.

$$Sim(z_i, z_j) = \frac{\sum_{k=1}^{V} p(w_k|z_i)p(w_k|z_j)}{\sqrt{\sum_{k=1}^{V} p(w_k|z_i)^2 \sum_{k=1}^{V} p(w_k|z_j)^2}} \qquad (2)$$

Figure 2 gives the visualization of topics and the relations between them. The red circle denotes the topic and the weighted edges between circles denote the relations between them. The larger the circle is, the hotter the topic. The thicker the line between topics, the stronger the relation they have. Users can get the overview and understand the main content of the news documents from the point of view of topics. They can know what topics are described in the news documents and know the relations between topics. As shown in Figure 2, we know that "金融体系"(Financial system) and "美国政府"(American Government) are two hot topics. We can find that the topic "美国经济"(American Economy) relates to the topic "经济衰退"(Economic recession), which may result in that "美国政府"(American Government) makes strategies to "刺激经济"(Jump-start Economy). We can also find that hot entities relevant to "金融监管"(Finance Supervision) are "格林斯潘"(Alan Greenspan), "蒋定之"(Jiang Dingzhi), "本.伯南克"(Ben Bernanke), etc.

Entities and Relations

In the NewsInsight system, we include four kinds of named entities. They are person (PER), location (LOC), organization (ORG) and time (TIM). The entity information is obtained using a language shallow processing tool. After obtaining the entity information, we focus on estimating the weight of relations between entities and extracting the relations' labels to represent the semantics of the relations.

Based on the learned topic model, we can get $p(e|z)$ which represents the probability of entity e given topic z. The weight of relations between two entities is defined as:

$$Sim(e_i, e_j) = \frac{\sum_{k=1}^{K} p(z_k|e_i)p(z_k|e_j)}{\sqrt{\sum_{k=1}^{K} p(z_k|e_i)^2 \sum_{k=1}^{K} p(z_k|e_j)^2}} \qquad (3)$$

$$\text{with } p(z_k|e_i) = \frac{p(e_i|z_k) \cdot p(z_k)}{p(e_i)} \qquad (4)$$

where $p(z_k)$ is the probability of the topic occurring in all news documents; $p(e_i)$ denotes the probability of the entity.

The other sub task is to determine the label of the relation. Our basic idea is to extract a subset of key words that are highly relevant to both of the entities to represent their relation. Formally, we denote this as $F(\vec{w}\|r(e_i, e_j))$, which is calculated as:

$$F(\vec{w}\|r(e_i, e_j)) = arg\max_{\vec{w}} \sum_{w_k \in \vec{w}} Sim(w_k, e_i) \cdot Sim(w_k, e_j) \qquad (5)$$

where $Sim(w_k, e_j)$ is calculated by Eq.3. The number of extracted words $\vec{w}$ for a relation $r(e_i, e_j)$ is tentatively set as 3. We can extract the label of relation between entities and topics in a similar way.

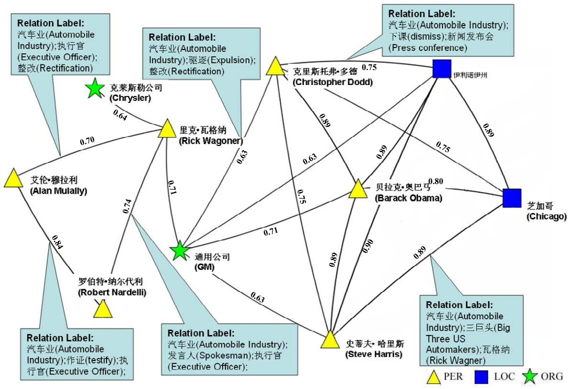

Relation Label:
汽车业(Automobile Industry);执行官(Executive Officer);整改(Rectification)
Relation Label:
汽车业(Automobile Industry);驱逐(Expulsion);整改(Rectification)
Relation Label:
汽车业(Automobile Industry);下课(dismiss);新闻发布会(Press conference)
Relation Label:
汽车业(Automobile Industry);作证(testify);执行官(Executive Officer);
Relation Label:
汽车业(Automobile Industry);发言人(Spokesman);执行官(Executive Officer);
Relation Label:
汽车业(Automobile Industry);三巨头(Big Three US Automakers);瓦格纳(Rick Wagner)
克莱斯勒公司 (Chrysler)
里克•瓦格纳 (Rick Wagoner)
克里斯托弗•多德 (Christopher Dodd)
伊利诺伊州
艾伦•穆拉利 (Alan Mulally)
贝拉克•奥巴马 (Barack Obama)
芝加哥 (Chicago)
通用公司 (GM)
罗伯特•纳尔代利 (Robert Nardelli)
史蒂夫• 哈里斯 (Steve Harris)
0.64
0.70
0.75
0.63
0.89
0.63
0.89
0.75
0.89
0.71
0.75
0.80
0.74
0.71
0.84
0.63
0.89
0.90
0.89
PER
LOC
ORG

Figure 3: Entities and relations.

Figure 3 illustrates the visualized result for entities and their relations, where yellow triangles, blue squares and green stars repre-

[3] http://finance.sina.com.cn/zt/index.html

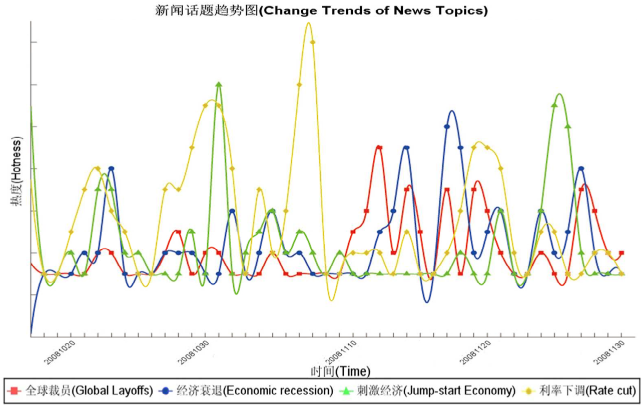


Figure 4: Change trend of example topics.

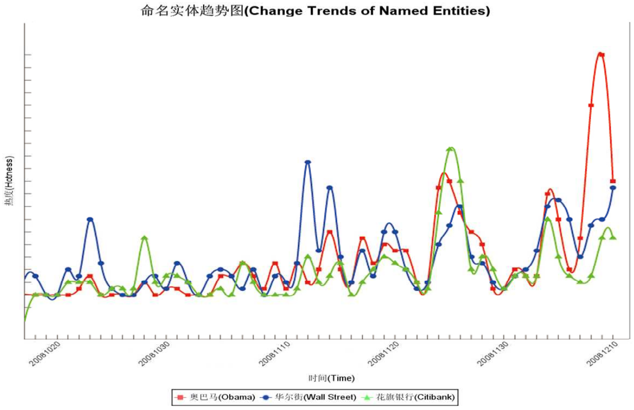


Figure 5: Change trend of example entities.

sent person (PER), location (LOC) and organization (ORG) respectively, and edges between them denote the relations between entities. For example, as shown in Figure 3, “里克.瓦格纳”(Rick Wagner), “艾伦.穆拉利”(Alan Mulally) and “罗伯特.纳尔代利”(Robert Nardelli) are three persons who have strong relations. They are related via “汽车业”(Automobile Industry) and “执行官”(Executive Officer). By further exploring the news documents with these relations, they are three CEOs for different international motor companies respectively (GM, Ford, and Chrysler).

Change Trend of Topics and Entities

Based on the discovered topic models, we conduct an analysis of the topic life cycles and entity dynamics. Figure 4 plots the occurring probability of four topics from the news data at different times. Figure 5 plots the occurring probability of three entities at different times.

From these two figures, we can see several interesting patterns. For example, as shown in Figure 4, in the mid of November of 2008, the news documents are mainly about the topic “经济衰退”(Economic recession). Soon, after one week, the news start to talk about the topic “刺激经济”(Jump-start Economy). Moreover, in the period of economic recession, there are many times that the topic “利率下调”(Rate cut) is highlighted in news documents, while with the continuing recession, not only do many governments set policies for stimulus, but also many big companies start to lay off employees.

Figure 5 also reflects some popular topic trends. For example, in the end of 2008, “奥巴马”(Obama) steadily became a star, especially when he won the president election. “花旗银行”(Citibank) also attracted a lot of attention during the period of economic downturn. As we know, the stock of “花旗银行”(Citibank) reduced to only 5% of its peak stock value. Over the whole period, “华尔街”(Wall Street) is always in the center of the discussion, which is clearly reflected in our analysis result (in Figure 5).

3. DEMONSTRATION PLAN

NewsInsight has been implemented in Java and is available at `http://keg.cs.tsinghua.edu.cn/software/NewsInsight`. The system has been used in the Xinhua agency, one of the biggest news publishers in China.

We will present our system thoroughly in the demonstration. Particularly, we will focus on the following aspects.

1. First, we will use a poster to give an overview of the system, including the motivation and major issues addressed in the system. We will introduce the architecture and the main features of the system.

2. Next, we will explain how we conduct the topic model for the news documents. In particular, we will describe in detail how we incorporate the information of entities and relations into the topic model. We will present how we identify the label of each topic and how to estimate the correlation between topics. We will further introduce how we characterize the relations between entities and how we capture the change trends of news topics and entities.

3. We will then demonstrate the visualization functions for each of the analysis results. We will show how a high-level overview of the news data is displayed to the user and how an in-depth view can be obtained if the user is interested in an entity (or topic).

4. Finally, we will share our thoughts about the strengths and the weaknesses of the system. We will further discuss plans for future work on the system.

4. ACKNOWLEDGMENTS

The work is supported by the Natural Science Foundation of China (60703059), National High-tech R&D Program (2009AA01Z138), Chinese National Key Foundation Research (2007CB310803), and Chinese Young Faculty Research Fund (20070003093).

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(993–1022), 2003.

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, pages 5228–5235, 2004.

[3] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.

[4] Y. H.-k. Liu Qun, Zhang Hua-Ping and C. Xue-Qi. Chinese lexical analysis using hierarchical hidden markov mode. *Chinese Journal of Computer Research and Development*, 41(8)(1421-1429), 2004.

[5] S.Jang and W.Woo. ubi-ucam: A unified context-aware application model. In *Proceedings of Context2005*, pages 178–189, 2003.

[6] C. Zhu, Weizhong. Chen. Storylines: Visual exploration and analysis in latent semantic spaces. *International Journal of Computers and Graphics. Special Issue on Visual Analytics*, 31(3)(338-349), 2007.