A Case Study of Multi-view News Exploration: Shanghai World Expo

Lu Lu, Lei Hou, Xiao Zhang, Zhichun Wang, Juanzi Li
Department of Computer Science and Technology, Tsinghua University
Haidian District, Beijing, China, 10084
{lulu, houlei, zhangxiao, zcwang, ljz}@keg.cs.tsinghua.edu.cn

ABSTRACT
In this paper, we introduce our system named multi-view news exploration system—NewsInsight. The system can help people to understand and explore news information from different views. We run the system on the 2010 Shanghai World Expo news data. Analysis result has been visualized and published on the web[1]. Through it, people can know important events and entities reported during 2010 Shanghai World Expo from multi-view, and read the news of their interest.

Keywords
News Mining, Topic Modeling, Entity Relation, Visualization

1. INTRODUCTION
With the rapid growth of news web sites, the internet has become one of the most convenient ways to access news information. According to the CNNIC's report, there are 80% of the 380 million internet users in China read news online in 2010[2]. Most news portals classify news into several categories and list news according to the happening time. People often spend much time in finding information of their interest from a large volume of news articles. Therefore, there is a great need for methods and tools to help users easily understand and explore news data. Some prior work on news mining and management either performs document based analysis such as clustering and classification news articles, or word based analysis such as identifying hot key words. In this paper, we present a news exploration system called NewsInsight [1]. NewsInsight recognizes three kinds of entities in news articles, namely Person, Organization and Location, and employs topic models to group related news. The system shows the facts about news topics and entities from multiple views to users. In this paper, we first introduce the architecture of NewsInsight and then present a case study on the 2011 Shanghai World Expo news.

2. OVERVIEW OF NEWSINSIGHT SYSTEM
As shown in Figure 1, NewsInsight system consists of four main components:

1. Data preprocessing: it includes two main sub-processes: tokenization and stop word filtering. After tokenization, we also identify three kinds of named entities in the news articles, including Person, Organization and Location.

2. Topic modeling: We use LDA[2] algorithm to clustering news articles into several topics, each topic is also labeled by the most related keywords in it.

3. Relation and Trend Analysis: We analyze three kinds of relations: relations between entities, relations between topics and relations between topic and entities. Based on the extracted topic model, we calculate the similarity between entities or topics as the weight of the relations, and label those relations with keywords which are highly relevant to both sides of the relations. Change trend is also measured with the number of news related with entities or topics through the timeline.

4. Web Client: NewsInsight presents the analyzing results of news in the form of web pages and charts. There are three kinds of views: overview view, entity and relation view, topic and entity trend view.

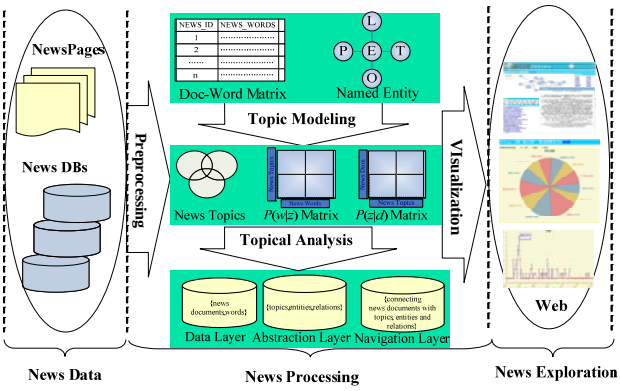

Figure 1. The architecture of News Explorer.

3. A CASE STUDY ON THE SHANGHAI WORLD EXPO NEWS
The news of the case study comes from the special issue of 2010 Shanghai World Expo on Chinese news portal site Sina.com. There are 9198 pieces of news from Dec. 9, 2009 to Nov. 15, 2010 in the specific news section. All the news dynamically reported information during the Expo. .For example, only in the opening day, there are 232 news reported in minutes. So, it is almost impossible for a user who wants to know and capture the main information occurring during Shanghai Expo. This is the focus of our paper. In this paper, we extracted the main topics from the news set, and quantitatively analysis the news from

WebSci '11, June 14-17, 2011, Koblenz, Germany.

1 http://keg.cs.tsinghua.edu.cn:8180/NI/

2 http://www.cnnic.net.cn/en/index/index.htm

multi-view including persons/organizations/places related to the topic.

3.1 News Topics Analysis

With the help of topic model analysis, we extract 20 topics from the news set and show the topics as a connected graph shown in Figure 2, in which each cycle represented a topic. The more important the topic is, the bigger the topic is. As can be seen, the most hot topic is 游客参观(Tourist and Visiting). Users can easily get a general understanding of all the topics and find the news they are interested in. In addition, when pointing the edge, a popup window will show the correlation between topics as well as the important common term in the topics. The right corner window shows the correlation between topic 游客参观(Tourist and Visiting) and 世博门票(Expo Tickets) with high correlation 0.51, which is reasonable.

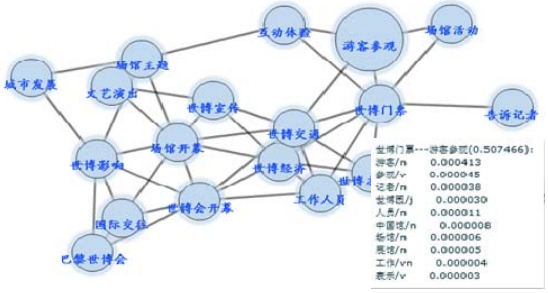

Figure 2. News topics and their relations.

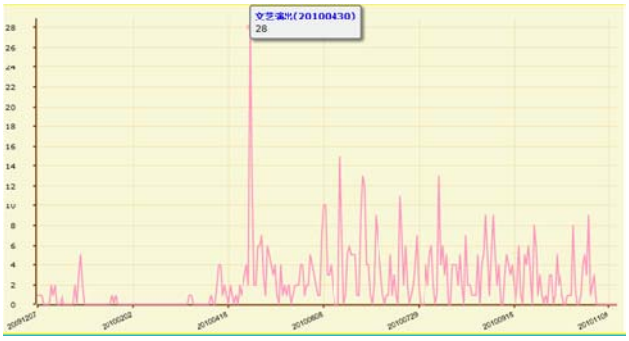

Figure 2. News topic trend

We also analyze the evolution of a given topic in the time span. Figure 3 displays the trend of topic 文艺演出 (Literary and Art Performance) by the number of news related to the topic on each date. As we can see from the figure, before the opening ceremony on Apr. 30, 2010, there is very little news about the topic. The opening ceremony date is the peak for this topic with 28 pieces of news. Then during the whole period of the Expo, although there is fluctuation on the news number, the trend is roughly stable. Other peaks on the figure correspond to some important activities. For example, Jun. 15, 2010, with 15 pieces of news, was the Ireland Pavilion Day, when they performed the famous tap dance and drew a lot of visitors and media. From the trend figure, users can easilyfind the important news when navigating news through topics.

3.2 News Entity Analysis

Our system explores three kinds of named entities, including persons, locations and organizations. In each topic, we list all the hot entities for users to see the details.

For example, in the 2010 World Expo news dataset, we can find person entity “董韵怡”(Yunyi Dong), location entity “英国”(United Kingdom) and organization entities“通用汽车馆”(GM Pavilion) and “伦敦零碳馆”(ZED Pavilion) in the topic “环保技术”(Environmental Technology) because the building of GM Pavilion and ZED Pavilion aims at environmental protection and Dong is CEO of GM in China. Similarly, we can find “Jianping Sun” who is the authority of Shanghai Transport and Port, some road name and “Shanghai Bus Cooperation” in the topic “Expo Transportation” which is obviously reasonable.

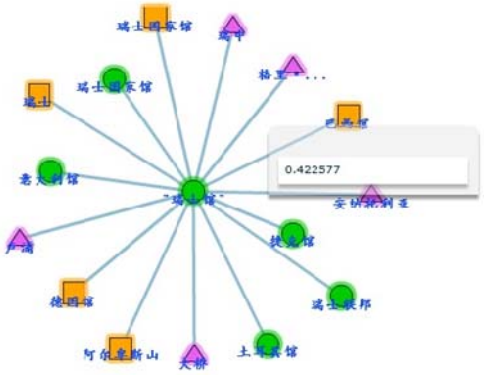

Figure 4. Entities’ relations in Topic

For each entity in one topic, we give out all of the related entities and corresponding news, for example, the organization entity“瑞士馆”(Swiss Pavilion) in the topic “场馆主题”(Pavilion Themes), as shown in figure 4. In figure 4, different kinds of named entities are represented with different shapes. The entity in the center is“瑞士馆”(Swiss Pavilion), and the others are entities related to it. There are person entity “格里·霍夫施泰特尔”(Gerry Hofstetter), location entity “阿尔卑斯山脉”(Alps) and organization entity “瑞士联邦”(Helvetica Confederation). User can put the mouse on the line linking two entities to get the value measuring their relation, for example, 04422577 between “瑞士馆”(Swiss Pavilion) and “格里·霍夫施泰特尔”(Gerry Hofstetter). If users want to know why they are linked, they can refer the corresponding news we provided.

We also provide two distribution functionalities in the entity analysis module. The first one is a pie chart displays the distribution of different kinds of entities under a given topic (as shown in Figure 5). The figure below shows the distribution of organizations in the topic “Pavilion Activity” in the 2010 Word Expo news set. As we can see, the hottest organizations in the topic include Spain Pavilion with weight 0.0306, Singapore Pavilion with weight 0.0217, South Africa Pavilion with weight 0.0183 and so on. The weight of each entity comes from the inference result of the topic model. The distribution indicates that the mentioned hottest organization held more reported activities than others involved in the 2010 World Expo, which is a correct conclusion according to the official statistics.

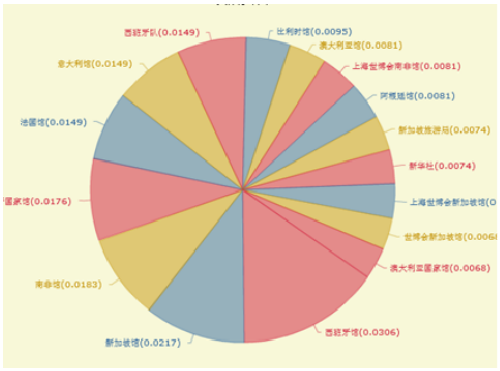

Figure 5. Distribution of Organizations in Topic

The other one is the temporal series of hottest entities in given topic (as shown in Figure 6). For example, the figure below displays the hottest organizations in the time line in topic "Tourist Visiting" in the 2010 World Expo news set. Each red point in the figure is the hottest organization on the given date. When the mouse pointed to the point, the popup window will show the detail of the organization. As in the figure, the popup window shows that the hottest organization of May 5th, 2010 is "Saudi Arabia Pavilion", which is mentioned in 9 news items from the set which exactly matches the circumstance that it is difficult to get a ticket for "Saudi Arabia Pavilion" at that time. From the time series, user can find how the focus of the topic changes with time.

Figure 5. Temporal Series of Entities in Topic

4. ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (No. 60973102, 61035004), the National Basic Research Program of China (973 Program) (No. 2007CB310803), it is also supported by IBM SUR joint project.

5. REFERENCES

[1] J. Li, J. Li, J. Tang, A Flexible Topic-driven Framework for News Exploration, SIGKDD2009

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993-1022, 2003