An Intelligent Search Platform for Business News

Hanchao Wang[1], Lili Zhou[1], Yu Zong[2], Lei Zhang[1], Enhong Chen[1], Xin Li[1], and Jun Chen[3]

[1] University of Science and Technology of China
[2] West Anhui University
[3] Xinhua News Agency
{hanchaow,zhoulili,stone}@mail.ustc.edu.cn, cheneh@ustc.edu.cn,
{nick.zongy,ustclixin}@gmail.com, chenjun2008@xinhua.org

Abstract. Living in a data driven world, the business news is very crucial for making economic decisions. To help decision makers obtain related business news quickly, two kinds of providers for business news, i.e., the search engine (e.g., Google News) and business portals (e.g., Reuters), are widely used. Though the keyword-based search engine is simple and easy to use, it has relatively low precision of the returned results and cannot directly provide news of particular business domains such as currency and real estate. In contrary, the portals can provide a variety of news of specific business domains, but it is difficult for users to browse since the front page looks so bloated and has many irrelevant ads. To solve the above problems, in this paper we propose and implement a platform named **I**ntelligent **S**earch **P**latform for **B**usiness **N**ews (ISPBN). This new platform not only combines the advantages of both search engine and portals, but also provides further analysis to discover the hidden relationships of different business news. To be specific, we incorporate automatic classification technology into the search platform to organize and retrieve business news in different domains. Furthermore, to fast guide users finding diversified and useful news, we construct a dynamic knowledge network graph to display the hidden relationships among news. Finally, we show the performance of our subsystems and present the final user interface of the proposed search platform.

Keywords: intelligent search, search engine, business news, web mining.

1 Introduction

With the high-speed development of the Internet, the information including all kinds of business news on the World Wide Web (WWW) is growing at an exponential rate. There is no doubt that the business news contains immense wealth and is very important to help people make decisions. However, people have to face the serious problem of information overload. Therefore, how to help users acquire valuable business news easily and quickly becomes a very vital problem.

F. Li et al. (Eds.): WAIM 2014, LNCS 8485, pp. 744–755, 2014.

Generally, there are two widely used providers for business news, namely the search engine (e.g., Google News [1]) and business portals (e.g., Reuters [2]). Most of us are now using search engine for information retrieval since it is very simple and easy to use. Users only need to input keywords, then they could acquire relevant results. However, too many returned results lead to relatively low precision and recall [1], and users have to spend a large amount of time on finding out useful information. More importantly, some users just care about news related to a particular field (e.g., currency, real estate) which the search engine cannot offer. Although some business portals provide the users with relatively professional and authoritative news of different business domains, there exists two drawbacks as follows. a) The home pages in these portals display all kinds of news related to different fields, which look bloated and huge. Therefore, it may confuse the users who are used to getting news by using search engine. b) These portals just simply display the news, and they cannot find out the hidden relationships of different business news, for example, the news about "housing price" may be related to news of "real estate control policy" or "building material industries".

In order to address the problems mentioned above, in this paper, we propose and implement a platform called **I**ntelligent **S**earch **P**latform for **B**usiness **N**ews (ISPBN). In this platform, we design and implement a vertical search engine system which incorporates automatic classification technology to organize and retrieve business news in different domains. In this way, our platform combines the advantages of both search engine and portals. The user can not only acquire news by keyword-based query, but also browse the news of specific fields by its category. To help users quickly understand business stories, we apply Name Entity Recognition (NER) techniques to recognize the entities (e.g., person names, time) in the clicked Web pages. Furthermore, we propose to construct dynamic knowledge network of news. Based on this knowledge network, we can discover the hidden relationships of news in different domains for news navigation. For example, if a user clicks a news story about "housing price", he can get the related news about "control policy" of "real estate" or news about "furniture industry". Our contributions can be summarized as follows:

- We propose and implement an Intelligent Search Platform for Business News (ISPBN). This novel platform combines the advantages of both search engine and business portals to help users acquire business news easily and quickly.
- We build a business thesaurus and implement an interactive management system used for optimizing our models such as classification model.
- We construct a dynamic knowledge network of news, which can be used to discover the hidden relationships of different business news. Based on this knowledge network, users can easily get the useful news they want.

The rest of this paper is organized as follows. We introduce the overview of ISPBN in Section 2 and describe the design of ISPBN in Section 3. We present the experimental results and final user interface in Section 4 and discuss the related work in Section 5. Finally, we conclude our work in Section 6.

[1] `https://news.google.com/`
[2] `http://www.reuters.com/`

2 Overview of ISPBN

We give an overview of ISPBN and introduce how each part of ISPBN works together in this section. As shown in Fig. 1, the architecture of our platform consists of four parts. 1) A vertical search engine including Intelligent Crawler,

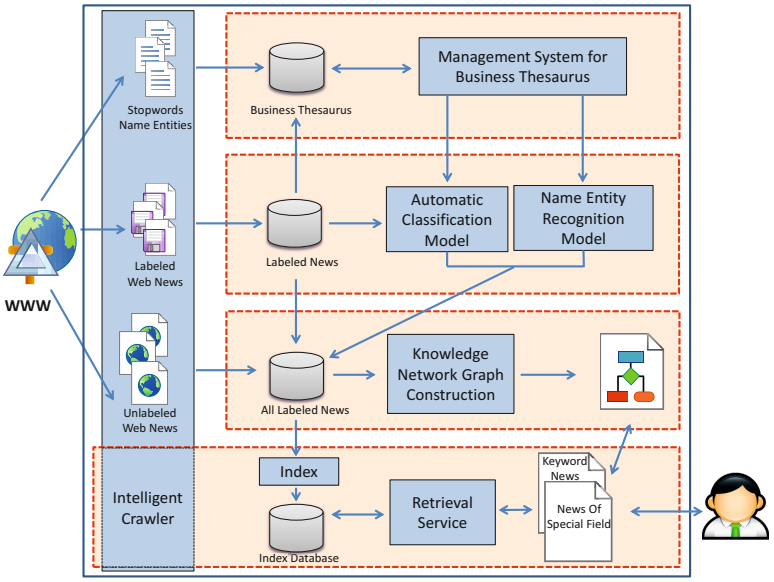

Fig. 1. The architecture of ISPBN

Index and Retrieval Service. 2) A Knowledge Network Graph Construction module. 3) Two models including Automatic Classification Model and Name Entity Recognition Model. 4) A Management System for Business Thesaurus. In the first part, we implement an intelligent crawler based on Nutch [3] and employ index and retrieval technologies based on Solr [4]. The crawler can get three kinds of data: the ordinarily unlabeled business news as the main part of index database, the specially labeled news used for training classification model and extracting feature words of each category, some famous name entities and stopwords. In the second part, we construct the knowledge network graph which can display the hidden relationships of different business news by analyzing all labeled news. In the third part, we integrate the search engine with two models, i.e., automatic classification model based on SVM [2] to provide the users with news of a specific field, and NER model to recognize important name entities of business news. Both of them are finished before indexing. In the fourth part, we design and implement a management system of business thesaurus for the sake of optimizing our models. Our thesaurus consists of stopwords, famous name entities and feature words which are extracted form the specially labeled news.

[3] `http://nutch.apache.org/`
[4] `http://lucene.apache.org/solr/`

3 Implementation of ISPBN

In this section, we describe the design and implementation of each system in our platform in detail. At first, we present the search engine system combined with automatic classification and NER techniques. Then we introduce how to build and manage the business thesaurus. Finally, we describle how to construct the dynamic knowledge network graph.

3.1 Vertical Search Engine

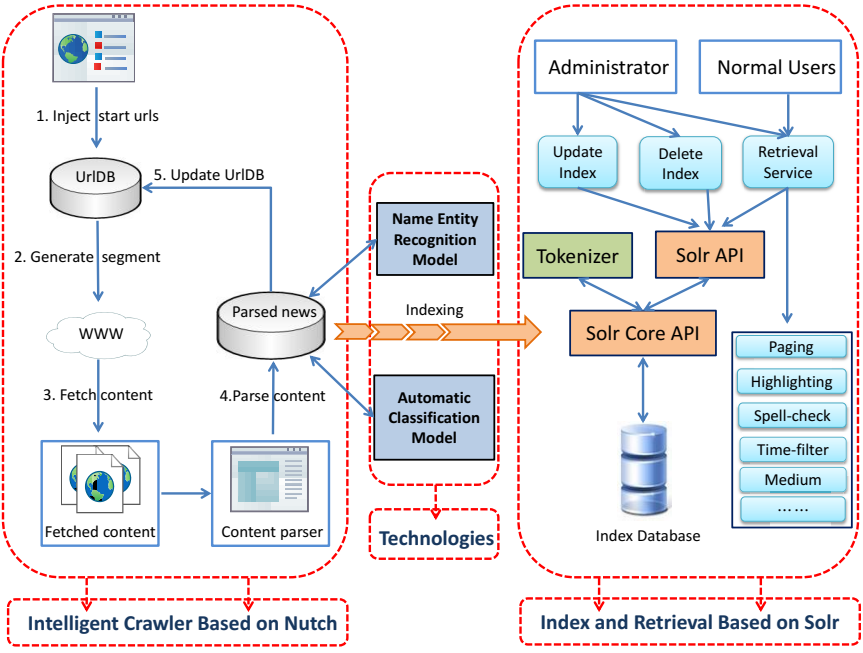

Fig. 2. The main components and techniques of the vertical search engine

The vertical search engine is the foundation of our platform. In this subsection, we not only introduce the vertical search engine, but also describe how to integrate automatic classification and NER technologies which are used for classifying unlabeled news and recognizing important name entities. Fig. 2 presents the main components and techniques of the vertical search engine. From Fig. 2, we can easily know that our vertical search engine is divided into three parts. We introduce each part as follows.

Intelligent Crawler Based on Nutch. As we all know, a search engine is formed with crawler, index and retrieval. The crawler is mainly responsible for acquiring a variety of data (e.g., web page, documents, videos, etc.). In our platform, we implement an intelligent crawler mainly used for crawling all kinds of business news. The implementation of our crawler is based on Nutch which is an open source web crawler software from the *Apache Lucene*^{TM} project. The

workflow of our crawler is divided into five steps. The crawler first injects start urls to Url Database (UrlDB). Then it generates segment which contains urls scheduled for fetching. After that, it fetches the content and parses the content by using content parser. Finally, it updates UrlDB with extracted urls. In order to build a vertical search engine in business field, we just crawl portals related to business. We periodically download business news from main business portals of China such as *sina.com* and *163.com*.

Automatic Classification and NER. Two kinds of business news, labeled and unlabeled, are crawled by our intelligent crawler. The labeled news is used to train our classification model. Before indexing, each unlabeled news story will be classified by its content into one of the eleven categories, namely, *real estate, textile industry, steel industry, chemical industry, finance and currencies, household appliance industry, furniture industry, construction and building material industries, employment, automobile industry and decoration industry*. Our automatic classifier is built on LIBSVM [3]. We use IK Analyzer [5] as the tokenizer. Each news story is automatically converted into a weight vector format. In addition, we take advantage of NER technology to recognise the person names, place names and organization names of each business news story, because these name entities (e.g., a company, an official spokesman, etc.) may be important to the users. In our platform, we adopt the Stanford Named Entity Recognizer [4] and the Chinese models which use distributional similarity clusters [6]. After processing by automatic classification and NER, we finally get all labeled news.

Index and Retrieval Based on Solr. The last part of search engine is index and retrieval. The implementation of this part is based on $Solr^{TM}$ which is a popular open source enterprise search platform from the *Apache* $Lucene^{TM}$ project. By calling Solr Core API, the parsed news will be stored in an index database. We provide management service of index database such as updating index or deleting index by calling Solr API just for the administrator. Any user including administrator and normal user, can enjoy retrieval service. We extend and enrich the retrieval functions of Solr. For instance, we implement functions like the results paging, highlighting both of title and content, spell check, time filter, the number of medium statistics, browsing news by category and so on.

3.2 Business Thesaurus

Each special field has its own feature words. When training a classification model, each news story is converted into a weight vector of the words. The weight can be calculated by using some classical and popular methods such as TF and IDF. Words in different fields have different weights. In a special field, we regard the words whose weight is higher than the general words as the feature words of this field. These feature words is vital for training models. Hence, we build a business thesaurus and implement a management system. In this subsection, we

[5] `http://code.google.com/p/ik-analyzer/`
[6] `http://nlp.stanford.edu/software/CRF-NER.shtml`

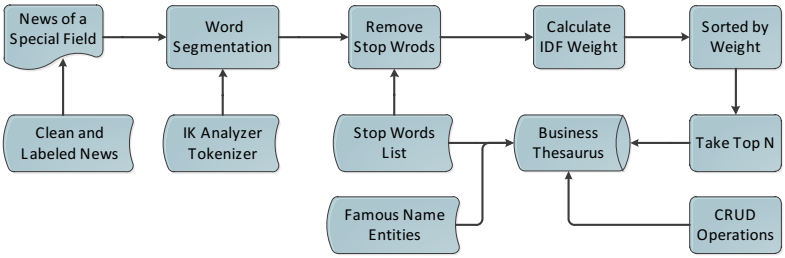

Fig. 3. The process of building and managing business thesaurus

introduce how to generate the feature words of each predefined field such as real estate, and we show how to extend and manage the thesaurus.

Fig. 3 illustrates the whole process of building and managing business thesaurus. Our thesaurus consists of three parts: the feature words of each predefined field, the stop words, the famous name entities (e.g., some person names, some company names, etc.). We take the news which is used for training classifier as the data source to extract feature words for each special field. Each news story must do Chinese word segmentation first. Then we remove stop words through the stop words list and calculate the weight using IDF. After that, we sort words in ascending order by their weight and take top N [7] words. Finally we put these words into the thesaurus database. We put stop words list into the database too, thus we can manage them by manual to optimize our model. In addition we use special crawler to acquire some famous name entities to assist our NER model. We implement the basic CRUD (create, read, update, delete) operations on this thesaurus database and build the management system. Experts could modify some inaccurate words or add new words in manual considering that there are a great number of noise in our corpus. The feedback of experts can be helpful for optimizing our models.

3.3 Dynamic Knowledge Network

News stories of different domains may have hidden relationships. For instance, a news story about housing price may have relationships with news stories of three major categories, namely upstream and downstream industries (e.g., construction and building material industries, furniture industry), industry policy such as real estate control policy, and macro-economy (e.g., stock, currency). If we can extract topics of news and find out the relationships of news according to their topics, we could guide and extend the users' interest. That's the main idea of our network. In this subsection, we mainly introduce how to construct and display the dynamic knowledge network graph. As shown in Fig. 4, the whole construction process consists of two main parts, namely offline topic extraction and online displaying.

[7] N is a user specified parameter. In our platform, it is set to 500 and it can be changed according to the size of the corpus.

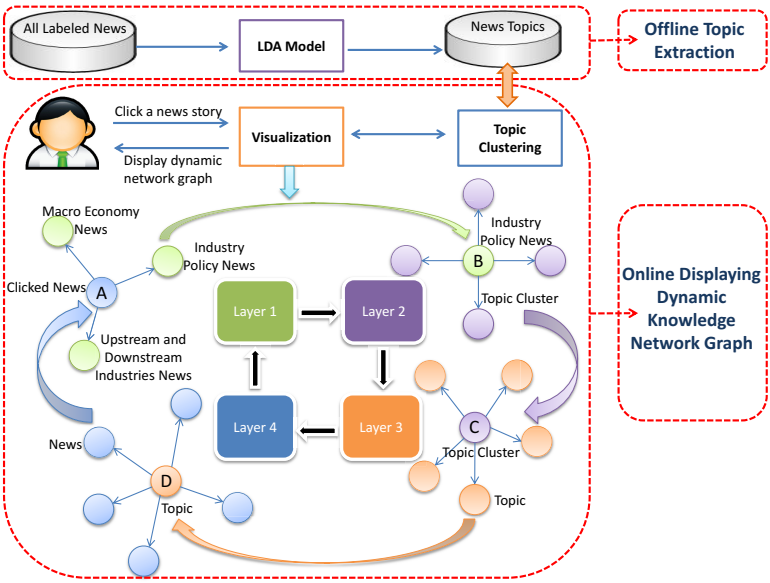

Fig. 4. The construction process of dynamic knowledge network graph. In this graph, (1) Node A represents a news story and is displayed by its title; (2) Node B represents a set of news stories and is displayed by its category; (3) Node C represents a cluster of topics namely a topic cluster and is displayed by a set of words, as we cluster the similar topics of a specific category; (4) Node D represents a specific topic and is displayed by a set of words.

Offline Topic Extraction. If a user clicks a news story, maybe he is interested in the news stories that have the same topic. We try to discover the hidden relationships of news according to their topics by using LDA (Latent Dirichlet Allocation) model [5]. LDA is a three-level hierarchical Bayesian model, in which each document of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is represented by a distribution over words. In our platform, we use JGibbLDA [8] to train LDA model and extract each news story's topics [6]. At the same time, the topics will be stored in a database. We design a table to store a news story's distribution over topics and another table to store a topic's distribution over words.

Online Displaying. When a user clicks a news story, as shown in Fig. 4, our platform will provide him with the hidden relationships through displaying the dynamic network graph. The network graph consists of four layers in due order. Each layer is the extension and supplement of the former layer. In the first layer, the center node denotes the news story that the user clicks. The other three nodes connected to the center node are a set of news stories which are related to the center news story and belong to three major categories including

[8] `http://jgibblda.sourceforge.net/`

macro economy, upstream and downstream industries, and industry policy. For instance, a user may care about housing price and he clicks a news story whose topics are about housing price. The first layer will extend the center news from three views to guide the user's interest. If the user is interested in news of the industry policy such as real estate control policy, he could click the industry policy node. Thus we step into the second layer, the center node is the industry policy and its each child node denotes a cluster of similar topics. It is the further arrangement and extension of the first layer. The user could click a topic cluster node that interests him. Thus we step into the third layer and each new node is a specific topic of the topics cluster. If the user clicks one of the unfolded nodes, the top K relevant news stories come out immediately as its children nodes. At the same time, we step into the fourth layer. The user could click and then browse the relevant news stories in detail. So far, we finish the construction process of dynamic knowledge network graph for the business news.

4 Experimental Results and Final User Interface

In this section, we evaluate the performance of the classifier and the impacts of our management system of thesaurus on the classifier. Then we present the final user interface of each system in the platform.

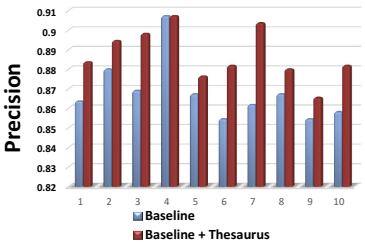

Fig. 5. The precision of the classifier

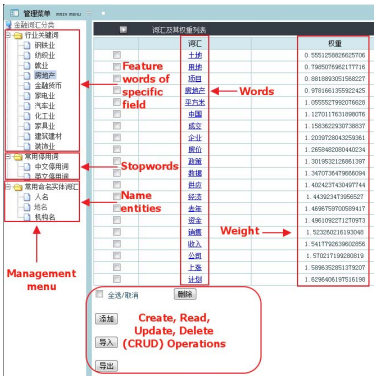

Fig. 6. The user interface of management system of business thesaurus

Classification Performance and the Management System Evaluation. To evaluate the performance of our classifier, a total of 5500 online news stories are collected by our crawler from five specific portals. These news stories are manually classified into eleven categories as mentioned in subsection 3.1, and each category has 500 news stories. We adopt 10-Cross-validation as the evaluation methodology and the average performance is reported. In order to evaluate the impacts of our management system of thesaurus on the classifier, we do another 10-Cross-validation experiments as a comparison. Fig. 5 shows the result of

our experiment. In the baseline method, we only use IK Analyzer to do Chinese Word Segmentation, while in our method we use IK Analyzer [9] combined with our thesaurus to do Chinese Word Segmentation. As shown, almost every set experiment in our method has the higher precision than the baseline method. It indicates that the thesaurus could optimize our model. The baseline has an overall precision of 86.84%, and in our method, the overall precision is improved to 88.73%. The user interface of our management system of our thesaurus is shown in Fig. 6.

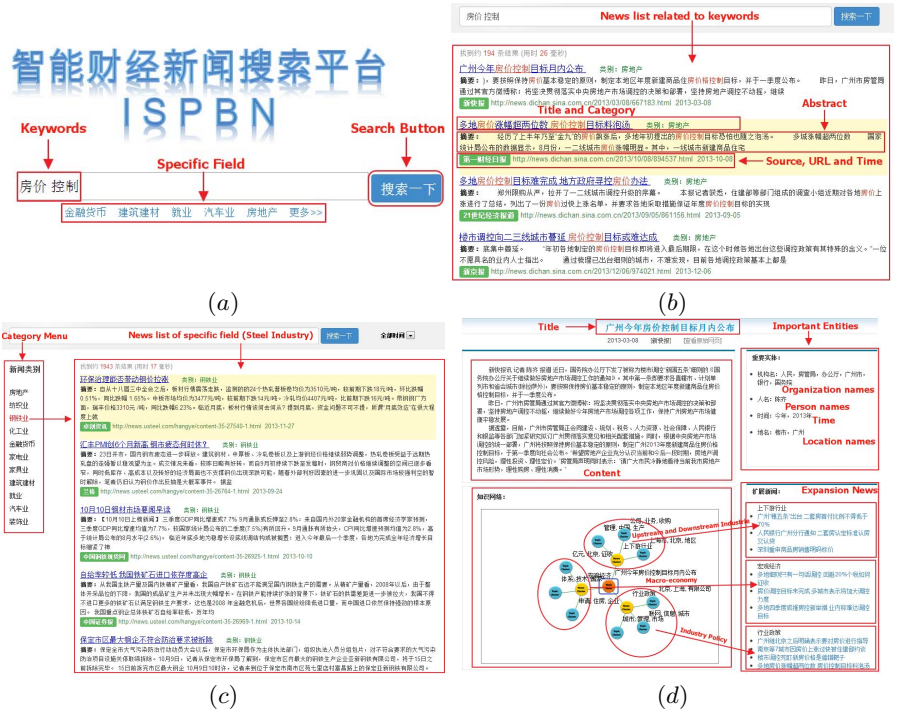

Fig. 7. The user interface of ISPBN: (a) The home page of ISPBN; (b) Get news by keyword-based search; (c) Browse news of specific fields; (d) Display the dynamic knowledge network graph, recommend name entities and extension news.

Final User Interface. Fig. 7 shows the user interface of our platform. From the home page in Fig. 7 (a), the user can not only acquire business news by inputting keywords as shown in Fig. 7(b), but also browse news of specific fields as shown in Fig. 7 (c). What'more, once the user clicks a news story, he will jump to another page which displays the dynamic knowledge network graph and recommends the important name entities and relevant news to the user in order to extend and guide their interest as shown in Fig. 7 (d).

[9] `http://code.google.com/p/ik-analyzer/`

5 Related Work

We present a brief survey of the relevant existing approaches both for intelligent search and web mining. In our platform we employ more web mining techniques aiming to make the search platform more intelligent and human-friendly.

5.1 Intelligent Search

In order to overcome the shortcomings of the traditional keyword based search engine, a great many researchers have focused on intelligent search. One of the hottest research topics is the Semantic Web [7]. There are a large number of related work on Semantic Web based intelligent search [8,9,10,11] and on Agent based intelligent search [12,13,14]. Specifically, [8] proposed a platform called QuestSemantics which provides automated ontology-based metadata creation and resource annotation based on a detailed ontological model of the domain, and [9] used the context-aware metadata which is stored in domain ontology. [10] presented the semantic web based search engine called SWISE which extracted metadata information of the pages and use the power of ontology. [11] investigated the semantic search performance of search engines by comparing with three keyword-based search engines. An agent based intelligent search framework for product information using ontology mapping was proposed in [12], while [13] presented the architecture of an agent-based intelligent search engine system for effective web mining. In this paper, we adopt web mining techniques to make our platform more intelligent. Our platform not only supports keyword-based search service, but also provides other useful service such browsing news by categories and knowledge graph navigation service.

5.2 Web Mining

Web mining is classified into three categories: Web content mining, Web structure mining and Web usage mining [15]. Web content mining aims to discover the useful information from web contents. There are two popular approaches used in web content mining, namely Agent based search approach and database approach [16,12]. There exists many popular web mining techniques for unstructured data such as information extraction techniques [17,18], topic detection and tracking techniques [19,20], summarization techniques [17], classification techniques [17,21], clustering techniques [20] and information visualization techniques [19,22]. For example, [17] proposed a system called Financial Information Digest System including three tasks (i.e., classification, information extraction, and information enquiry). [21] presented a Theme-Based News Retrieval System which incorporated a classification framework based on Support Vector Machines. We also adopted SVM in our platform, however we built a management system of business thesaurus to further improve the accuracy of the classification model. In addition, we proposed to construct a novel knowledge network which can be used to mine the hidden relationships of different business news. [19] also tried to understand the mutual relationships between information flows and social activity. Note that [19] aimed to explain abnormal financial market volatility, while we tried to guide and extend the users' interest.

6 Conclusion and Future Work

In this paper, we proposed and implemented an Intelligent Search Platform for Business News (ISPBN). This platform could satisfy enterprise search requirement because it combines the features of two open source projects of Apache (namely Nutch and Solr). Moreover, we integrated classification techniques into the search engine, so as to make it easy for users to acquire and browse news stories in different domains as well as by using keyword query. In the mean time, we built a management system of business thesaurus which could be used for optimizing models. In addition, to help users quickly understand business stories, we also incorporated name entity recognition techniques into the search platform to recognize the name entities (e.g., person names, organization names) in the clicked Web pages. Furthermore, we proposed to construct dynamic knowledge network graph based on business news, which can be used to find out the hidden relationships of news in different domains for news navigation. Based on this platform, users can easily and quickly get the information they need.

For the future work, we try to build the users' personal profiles and implement a personal recommendation system according to the browsing and searching history log. We hope that our platform could provide the users with more accurate, personalized and intelligent service to help them make decisions.

Acknowledgments. This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the National High Technology Research and Development Program of China (Grant No. 2014AA015203), the Science and Technology Development of Anhui Province, China (Grants No. 13Z02008-5 and 1301022064), the International Science & Technology Cooperation Plan of Anhui Province (Grant No. 1303063008), the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2012BAH17B03) and the Nature Science Research of Anhui (Grant No. 1208085MF 95).

References

1. Chakrabarti, S.: Data mining for hypertext: A tutorial survey. ACM SIGKDD Explorations Newsletter 1(2), 1–11 (2000)
2. Vapnik, V.: The nature of statistical learning theory. Springer (2000)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363–370. Association for Computational Linguistics (2005)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)

6. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, pp. 91–100. ACM (2008)
7. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific American 284(5), 28–37 (2001)
8. Tamma, V.: Semantic web support for intelligent search and retrieval of business knowledge. IEEE Intelligent Systems 25(1), 84–88 (2010)
9. Khattak, A.M., Mustafa, J., Ahmed, N., Latif, K., Khan, S.: Intelligent search in digital documents. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2008, vol. 1, pp. 558–561. IEEE (2008)
10. Shaikh, F., Siddiqui, U.A., Shahzadi, I., Jami, S.I., Shaikh, Z.A.: Swise: Semantic web based intelligent search engine. In: 2010 International Conference on Information and Emerging Technologies (ICIET), pp. 1–5. IEEE (2010)
11. Tumer, D., Shah, M.A., Bitirim, Y.: An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia. In: Fourth International Conference on Internet Monitoring and Protection, ICIMP 2009, pp. 51–55. IEEE (2009)
12. Inamdar, S., Shinde, G.: An agent based intelligent search engine system for web mining. Research, Reflections and Innovations in Integrating ICT in Education (2008)
13. Kim, W., Choi, D.W., Park, S.: Agent based intelligent search framework for product information using ontology mapping. Journal of Intelligent Information Systems 30(3), 227–247 (2008)
14. Hai-long, C.: Design and realization of intelligent search engine based on multi-agents [j]. Journal of Harbin University of Commerce (Natural Sciences Edition) 2, 016 (2009)
15. Al-Azmi, A.A.R.: Data, text, and web mining for business intelligence: A survey. International Journal of Data Mining & Knowledge Management Process 3(2) (2013)
16. Srividya, M., Anandhi, D., Ahmed, M.I.: Web mining and its categories–a survey. International Journal of Engineering and Computer Science, IJECS 2(4), 1338–1345 (2013)
17. Lam, W., Ho, K.S.: Fids: an intelligent financial web news articles digest system. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 31(6), 753–762 (2001)
18. Domenech, J.: An intelligent system for retrieving economic information from corporate websites. In: Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 573–578. IEEE Computer Society (2012)
19. Hisano, R., Sornette, D., Mizuno, T., Ohnishi, T., Watanabe, T.: High quality topic extraction from business news explains abnormal financial market volatility. PloS One 8(6), e64846 (2013)
20. Dai, X.Y., Chen, Q.C., Wang, X.L., Xu, J.: Online topic detection and tracking of financial news based on hierarchical clustering. In: 2010 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 6, pp. 3341–3346. IEEE (2010)
21. Maria, N., Silva, M.J.: Theme-based retrieval of web news. In: Suciu, D., Vossen, G. (eds.) WebDB 2000. LNCS, vol. 1997, pp. 26–37. Springer, Heidelberg (2001)
22. Gupta, V., Lehal, G.S.: A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence 1(1), 60–76 (2009)