Web Information Mining and Decision Support Platform for the Modern Service Industry

Binyang Li[1,2], Lanjun Zhou[2,3], Zhongyu Wei[2,3], Kam-fai Wong[2,3,4], Ruifeng Xu[5], Yunqing Xia[6]

[1] Dept. of Information Science & Technology, University of International Relations, China
[2]Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[3]MoE Key Laboratory of High Confidence Software Technologies, China
[4] Shenzhen Research Institute, The Chinese University of Hong Kong
[5]Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[6]Department of Computer Science & Technology, TNList, Tsinghua University, China
{byli,ljzhou,zywei,kfwong}@se.cuhk.edu.hk

Abstract

This demonstration presents an intelligent information platform MODEST. MODEST will provide enterprises with the services of retrieving news from websites, extracting commercial information, exploring customers' opinions, and analyzing collaborative/competitive social networks. In this way, enterprises can improve the competitive abilities and facilitate potential collaboration activities. At the meanwhile, MODEST can also help governments to acquire information about one single company or the entire board timely, and make prompt strategies for better support. Currently, MODEST is applied to the pillar industries of Hong Kong, including innovative finance, modem logistics, information technology, etc.

1 Introduction

With the rapid development of Web 2.0, the amount of information is exploding. There are millions of events towards companies and billions of opinions on products generated every day (Liu, 2012). Such enormous information cannot only facilitate companies to improve their competitive abilities, but also help government to make prompt decisions for better support or timely monitor, e.g. effective risk management. For this reason, there is a growing demand of Web information mining and intelligent decision support services for the industries. Such services are collectively referred as modern service, which includes the following requirements:

(1) To efficiently retrieve relevant information from the websites;
(2) To accurately determine the latest business news and trends of the company;
(3) To identify and analyze customers' opinions towards the company;
(4) To explore the collaborative and competitive relationship with other companies;
(5) To leverage the knowledge mined from the business news and company social network for decision support.

In this demonstration, we will present a Web information mining and decision support platform, MODEST[1]. The objective of MODEST is to provide modern services for both enterprises and government, including collecting Web information, making deep analysis, and providing supporting decision. The innovation of MODEST is focusing on deep analysis which incorporates the following functions:

- Topic detection and tracking function is to cluster the hot events and capture the relationship between the relevant events based on the collected data from websites (*event* also referred as *topic* in this paper). In order to realize this function, Web mining techniques are adopted, e.g. topic clustering, heuristics algorithms, etc.
- The second function is to identify and analyze customers' opinions about the company. Opinion mining technology (Zhou et al., 2010) is adopted to determine the polarity of those news, which can help the company timely and appropriately adjust the policy to strengthen the dominant position or avoid risks.

[1] This work is supported by the Innovation and Technology Fund of Hong Kong SAR.

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 97–102, Baltimore, Maryland USA, June 23-24, 2014.

- The third function is to explore and analyze social network based on the company centric. We utilize social network analysis (SNA) technology (Xia et al., 2010) to discover the relationships, and we further analyze the content in fine-grained granularity to identify its potential partners or competitors.

With the help of MODEST, the companies can acquire modern service-related information, and timely adjust corporate policies and marketing plan ahead. Hence, the ability of information acquisition and the competitiveness of the enterprises can be improved accordingly.

In this paper, we will use a practical example to illustrate our platform and evaluate the performance of main functions.

The rest of this paper is organized as follows. Section 2 will introduce the system description as well as the main functions implementation. The practical case study will be illustrated in Section 3. The performance of MODEST will be evaluated in Section 4.Finally, this paper will be concluded in Section 5.

2 System Description

In this section, we first outline the system architecture of MODEST, and then describe the implementation of the main functionality in detail.

2.1 Architecture and Workflow

The MODEST system consists of three modules: data acquisition, data analysis, and result display. The system architecture is shown in Figure 1.

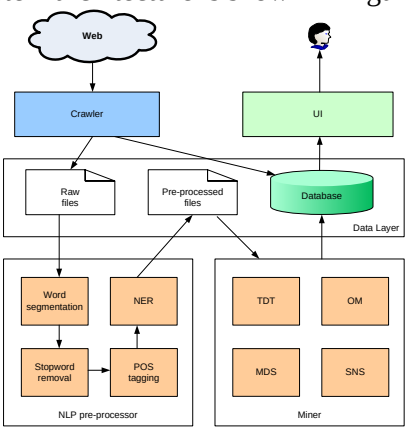


Figure 1: System architecture. (The module in blue is data acquisition, the module in orange is data analysis, and the module in light green is result display)

(1) The core technique in the data acquisition module is the crawler, which is developed to collect raw data from websites, e.g. news portals, blogosphere. Then the system parse the raw web pages and extract information to store in the local database for further processing.

(2) The data analysis module can be divided into two parts:

- NLP pre-processor: utilizes NLP (natural language processing) techniques and some toolkits to perform the pre-processing on the raw data in (1), including word segmentation, part-of-speech (POS) tagging[1], stopword removal, and named entity recognition (NER)[2]. We then create knowledgebase for individual industry, such as domain-specific sentiment word lexicon, name entity collection, and so on.
- Miner：makes use of data mining techniques to realize four functions, topic detection and tracking (TDT), multi-document summarization[3] (MDS), social network analysis (SNA), and opinion mining (OM). The results of data analysis are also stored in the database.

(3) The result display module read out the analysis results from the database and display them to users in the form of plain text, charts, figures, as well as video.

2.2 Function Implementation

Since the innovation of MODEST is focusing on the module of data analysis, we will describe its main functions in detail, including topic detection and tracking, opinion mining, and social networks analysis.

2.2.1 Topic Detection and Tracking

The TDT function targets on detecting and tracking the hot topics for each individual company. Given a period of data collected from websites, there are various discussions about the company. In order to extract these topics, clustering methods (Viermetz et al., 2007 and Yoon et al., 2009) are implemented to explore the topics. Note that during the period of data collection, different topics with respect to the same company may have relations. We, therefore, utilize hierarchical clustering methods[4]to capture the potential relations.

Due to the large amount of data, it is impossible to view all the topics at a snapshot. MODEST utilizes topic tracking technique (Wang et al., 2008) to identify related stories with a stream of

[1] www.ictclas.org
[2]http://ir.hit.edu.cn/demo/ltp
[3]http://libots.sourceforge.net/
[4]http://dragon.ischool.drexel.edu/

media. It is convenient for the users to see the latest information about the company.

In summary, TDT function provides the services of detecting and tracking the latest and emergent topics, analyzing the relationships of topics on the dynamics of the company. It meets the aforementioned demand, "to accurately grasp the latest business news and trends of the company".

2.2.2 Opinion Mining

The objective of OM function is to discover opinions towards a company and classify the opinions into positive, negative, or neutral.

The opinion mining function is redesigned based on our own opinion mining engine (Zhou et al., 2010). It separates opinion identification and polarity classification into two stages.

Given a set of documents that are relevant to the company, we first split the documents into sentences, and then identify whether the sentence is opinionated or not. We extract the features shown in Table 1 for opinion identification. (Zhou et al., 2010)

Table 1: Features adopted in the opinionated sentence classifier

Punctuation level features
The presence of direct quote punctuation "“" and "”"
The presence of other punctuations: "?" and "!"
Word-Level and entity-level features
The presence of known opinion operators
The percentage of known opinion word in sentence
Presence of a named entity
Presence of pronoun
Presence of known opinion indicators
Presence of known degree adverbs
Presence of known conjunctions
Bi-gram features
Named entities + opinion operators
Pronouns + opinion operators
Nouns or named entities + opinion words
Pronouns + opinion words
Opinion words (adjective) + opinion words(noun)
Degree adverbs + opinion words
Degree adverbs + opinion operators

These features are then combined using a radial basis function (RBF) kernel and a support vector machine (SVM) classifier (Drucker et al., 1997) is trained based on the NTCIR 8training data for opinion identification (Kando, 2010).

For those opinionated sentences, we then classify them into positive, negative, or neutral. In addition to the features shown in Table 1, we incorporate features of s-VSM (Sentiment Vector Space Model) (Xia et al., 2008) to enhance the performance. The principles of the s-VSM are listed as follows: (1) Only sentiment-related words are used to produce sentiment features for the s-VSM. (2) The sentiment words are appropriately disambiguated with the neighboring negations and modifiers. (3) Negations and modifiers are included in the s-VSM to reflect the functions of inversing, strengthening and weakening. Sentiment unit is the appropriate element complying with the above principles. (Zhou et al., 2010)

In addition to polarity classification, opinion holder and target are also recognized in OM function for further identifying the relationship that two companies have, e.g. collaborative or competitive. Both of the dependency parser and the semantic role labeling[1] (SRL) tool are incorporated to identify the semantic roles of each chunk based on verbs in the sentence.

The OM function provides the company with services of analyzing the social sentimental feedback on the dynamics of the company. It meets the aforementioned demand, "to identify and analyze customers' opinions towards the company".

2.2.3 Social Network Analysis

SNA function aims at producing the commercial network of companies that are hidden within the articles.

To achieve this goal, we maintain two lexicons, the commercial named entity lexicon and commercial relation lexicon. Commercial named entity are firstly located within the text and then recorded in the commercial entity lexicon in the pre-processor NER. Commercial relation lexicon record the articles/documents that involve the commercial relations. Note that the commercial relation lexicon (Table 2) is manually compiled. In this work, we consider only two general commercial relations, namely cooperation and competition.

Table 2: Statistics on relation lexicon.

Type	Amount	Examples
Competition	20	挑战(challenge), 竞争(compete), 对手(opponent)
Collaboration	18	协作(collaborate),协同(coordinate), 合作(cooperate)

SNA function produces the social network of a centric company, which can provide the compa-

[1]http://ir.hit.edu.cn/demo/ltp

ny with the impact analysis and decision-making chain tracking. It meets the aforementioned demand, "to explore the collaborative and competitive relationship between companies".

3 Practical Example

In this section, we use a case study to illustrate our system and further evaluate the performance of the main functions with respect to those companies. Due to the limited space, we just illustrate the main functions of topic detection, opinion mining and social network analysis.

3.1 Topic Detection and Opinion Mining

Figure 2(a) showed the results of topic detection and opinion mining functions for a Hong Kong local financial company *Sun Hung Kai Properties* (新鴻基地產). On top of the figure are the results of topic detection and tracking function. Multi-document summary of the latest news is provided for the company and more news with the similar topics can be found by pressing the button "更多" (*more*). Since there are a lot of duplicates of a piece of news on the websites, the summary is a direct way to acquire the recent news, which can improve the effectiveness of the company.

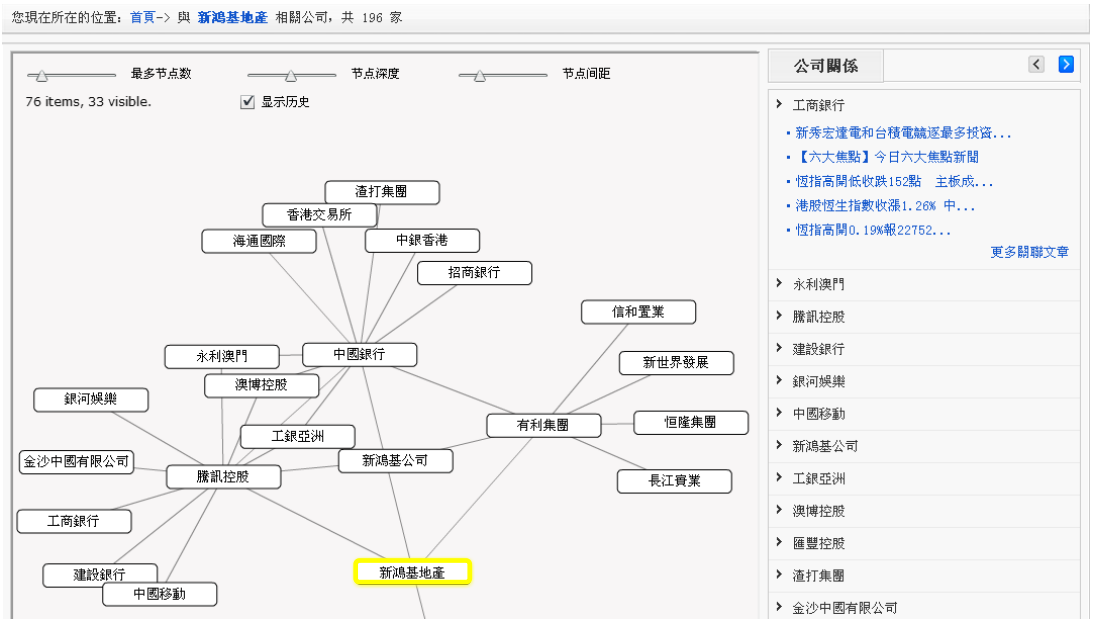
The results of opinion mining function are shown at the bottom of Figure 2(a), where the green line indicates negative while the red line indicates positive. In order to give a dynamic insight of public opinions, we provide the amount changes of positive and negative articles with time variant. This is very helpful for the company to capture the feedback of their marketing policies. As shown in Figure 2(a), there were 14 negative articles (負面信息) on Oct. 29, 2012, which achieved negative peak within the 6 months. The users would probably read those 14 articles and adjust the company strategy accordingly.

3.2 Social Network Analysis

Figure 2(b) shows the social network based on the centric company in yellow, *Sun Hung Kai Properties* (新鴻基地產). We only list the half of the connected companies with collaborative relationship from *Sun Hung Kai Properties*, and remove the competitive ones due to limited space. The thickness of the line indicates the strength of the collaboration between the two companies. The social network can explore the potential partners/competitors of a company. Furthermore, users are allowed to adjust the depth and set the nodes count of the network. The above analysis can provide a richer insight in to a company.

In the following section, we will make experiments to investigate the performance of the above functions.

(a) Topic detection and opinion mining of *Sun Hung Kai Properties* (新鴻基地產). (For convenience, we translate the texts on the button in English)



(b)Social network of *Sun Hung Kai Properties* (新鴻基地產). (The rectangle in yellow is the centric)

Figure 2: Screenshot of the MODEST system.

4 Experiment and Result

In our evaluation, the experiments were made based on 17692 articles collected from 52 Hong Kong websites during 6 months (1/7/2012~ 31/12/2012). We investigate the performance of MODEST based on the standard metrics proposed by NIST[1], including precision, recall, and F-score.

Precision (*P*) is the fraction of detected articles (*U*) that are relevantto the topic (*N*).

$$P = \frac{N}{U} \times 100\%$$

Recall (*R*) is the fraction of the articles (*T*) that are relevant to the topic that are successfully detected (*N*).

$$R = \frac{N}{T} \times 100\%$$

Usually, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. Therefore, precision and recall scores are not discussed in isolation. Instead, F-Score (*F*) is proposed to combine precision and recall, which is the harmonic meanof precision and recall.

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} \times 100\% = \frac{2 \times P \times R}{P + R} \times 100\%$$

4.1 Topic Detection and Tracking

We first assess the performance of the topic detection function. The data is divided into 6 parts according to the time. For different companies, the amount of articles vary a lot. Therefore, we calculate the metrics for each individual dataset, and then compute the weighted mean value. The experimental results are shown in Table 3.

Table 3: Experimental results on topic detection.

Dataset	Recall	Precision	F-Score
1/7/12-31/7/12	85.71%	89.52%	85.38%
1/8/12-31/8/12	93.10%	93.68%	92.49%
1/9/12-30/9/12	76.50%	83.13%	76.56%
1/10/12-31/10/12	83.32%	88.53%	85.84%
1/11/12-30/11/12	86.11%	89.94%	87.98%
1/12/12-31/12/12	84.26%	87.65%	85.92%
Average	85.13%	88.78%	85.69%

From the experimental results, we can find that the average F-Score is about 85.69%.The dataset in the second row achieves the best performance while the dataset in the third only get 76.56% in F-Score. It is because that the amount of articles is smaller than the others and the recall value is very low. As far as we know, the best run of topic detection in (Allan et al., 2007) achieved 84%. The performance of topic detection in MODEST is comparable.

4.2 Opinion Mining

We then evaluate the performance of opinion mining function. We manually annotated 1568 articles, which is further divided into 8 datasets randomly. Precision, recall, and F-score are also used as the metrics for the evaluation. The experimental results are shown in Table 4.

[1]http://trec.nist.gov/

Table 4: Experimental results on opinion mining.

Dataset	Size	Precision	Recall	F-Score
dataset-1	200	76.57%	78.26%	76.57%
dataset-2	200	83.55%	89.64%	86.07%
dataset-3	200	69.12%	69.80%	69.44%
dataset-4	200	77.13%	75.40%	75.67%
dataset-5	200	76.21%	77.65%	76.74%
dataset-6	200	63.76%	66.22%	64.49%
dataset-7	200	78.56%	78.41%	78.43%
dataset-8	168	65.72%	65.15%	65.32%
Average	196	73.83%	75.07%	74.09%

From Table 4, we can find that the average F-Score can reach 74.09%. Note that the opinion mining engine of MODEST is the implementation of (Zhou et al., 2010), which achieved the best run in NTCIR. However, the engine is trained on NTCIR corpus, which consists of articles of general domain, while the test set focuses on the financial domain. We further train our engine on the data from the financial domain and the average F-Score improves to over 80%.

5 Conclusions

This demonstration presents an intelligent information platform designed to mine Web information and provide decisions for modern service, MODEST. MODEST can provide the services of retrieving news from websites, extracting commercial information, exploring customers' opinions about a given company, and analyzing its collaborative/competitive social networks. Both enterprises and government are the target customers. For enterprise, MODEST can improve the competitive abilities and facilitate potential collaboration. For government, MODEST can collect information about the entire industry, and make prompt strategies for better support.

In this paper, we first introduce the system architecture design and the main functions implementation, including topic detection and tracking, opinion mining, and social network analysis. Then a case study is given to illustrate the functions of MODEST. In order to evaluate the performance of MODEST, we also conduct the experiments based on the data from 52 Hong Kong websites, and the results show the effectiveness of the above functions.

In the future, MODEST will be improved in two directions:

- Extend to other languages, e.g. English, Simplified Chinese, etc.
- Enhance the compatibility to implement on mobile device.

The demo of MODEST and the related toolkits can be found on the homepage: http://sepc111.se.cuhk.edu.hk:8080/adcom_hk/

Acknowledgements

This research is partially supported by General Research Fund of Hong Kong (417112), Shenzhen Fundamental Research Program (JCYJ20130401172046450, JCYJ20120613152557576), KTO(TBF1ENG007), National Natural Science Foundation of China (61203378, 61370165), and Shenzhen International Cooperation Funding (GJHZ20120613110641217).

References:

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic Detection and Tracking Pilot Study: Final Report. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vpnik. 1997. Support Vector Regression Machines. Proceedings of Advances in Neural Information Processing Systems, pp. 155-161.

Noriko Kando.2010. Overview of the Eighth NTCIR Workshop. Proceedings of NTCIR-8 Workshop.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. Proceedings of Synthesis Lectures on Human Language Technologies, pp. 1-167.

Maximilian Viermetz, and Michal Skubacz. 2007. Using Topic Discovery to Segment Large Communication Graphs for Social Network Analysis. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 95-99.

Canhui Wang, Min Zhang, Liyun Ru, and Shaoping Ma. 2008. Automatic Online News Topic Ranking Using Media Focus and User Attention based on Aging Theory. Proceedings of the Conference on Information and Knowledge Management.

Yunqing Xia, Nianxing Ji, Weifeng Su, and Yi Liu. 2010. Mining Commercial Networks from Online Financial News. Proceedings of the IEEE International Conference on E-Business Engineering, pp. 17-23.

Ruifeng Xu, Kam-fai Wong, and Yunqing Xia. 2008. Coarse-Fine Opinion Mining-WIA in NTCIR-7 MOAT Task. In NTCIR-7 Workshop, pp. 307-313.

Seok-Ho Yoon, Jung-Hwan Shin, Sang-Wook Kim, and Sunju Park. 2009. Extraction of a Latent Blog Community based on Subject. Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 1529-1532.

Lanjun Zhou, Yunqing Xia, Binyang Li, and Kam-fai Wong. 2010. WIA-Opinmine System in NTCIR-8 MOAT Evaluation. Proceedings of NTCIR-8 Workshop Meeting, pp. 286-292.