Advanced Science and Technology Letters
Vol.29 (DTA 2013), pp.372-378
http://dx.doi.org/10.14257/astl.2013.29.77

Design Consideration on a Real-time System for Collecting Intelligence from the Web

Jie Zhao[1], Peiquan Jin[2]

[1]School of Business, Anhui University, 230601, Hefei, China
[2]School of Computer Science and Technology,
University of Science and Technology of China, 2300027, Hefei, China
zj_teacher@126.com

Abstract. In this paper, we propose a real-time system to collect intelligence from the Web. We first divide the intelligence in the Web into three types, namely single-source intelligence, multi-source intelligence, and strategic intelligence, and then the system framework to collect those types of intelligence is discussed.

Keywords: Intelligence, Web, Information Collection

1 Introduction

Web as a rich data source differs from other types of data sources. Generally, Web data has three characters: large-volume, public, and interactive. Presently there are hundreds of millions of Web sites, which produce large amount of Web data. On the other side, compared with other types of data such as finance data, Web data is free for access. Furthermore, with the rapid development of Web 2.0, Web data becomes more and more interactive. Blog, microblog, RSS, Wikipedia, and other Web 2.0 services enable us to act as both data consumers and data producers.

There is a lot of valuable information, or in other words, intelligence, hidden in the Web, including military intelligence, economic intelligence, political intelligence, and so on [1-3]. For example, it is very easy to find information about a newly founded company in the Web. Therefore, if we can build a system to collect intelligence in the Web in a real time way, it will bring new values for both governments and companies.

2 Types of Intelligence in the Web

The intelligence hidden in the Web can basically be divided into three categories, namely single-source intelligence, multi-source intelligence, and strategic intelligence.

ISSN: 2287-1233 ASTL

2.1 Single-Source Intelligence

The single-source intelligence is from one Web site. It usually contains fresh news information, but is with low credibility. The single-source intelligence can be regarded as the source of new intelligence, which forms the base of multi-source intelligence and strategic intelligence.

Fig.1 shows the process to collect single-source intelligence from the Web. The input of the algorithm is an intelligence seed. It utilizes the search engines to find sensitive sources of intelligence and monitor those sources to out sensitive intelligence. The sensitive sources discovery can be performed using some natural information processing techniques, such as word segmentation.

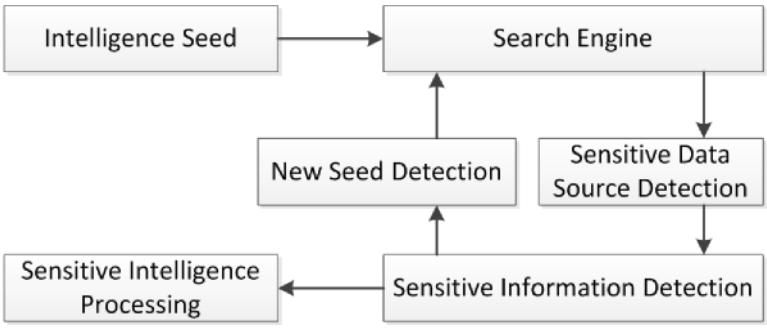

Fig. 1. The process of collecting single-source intelligence from the Web

2.2 Multi-Source Intelligence

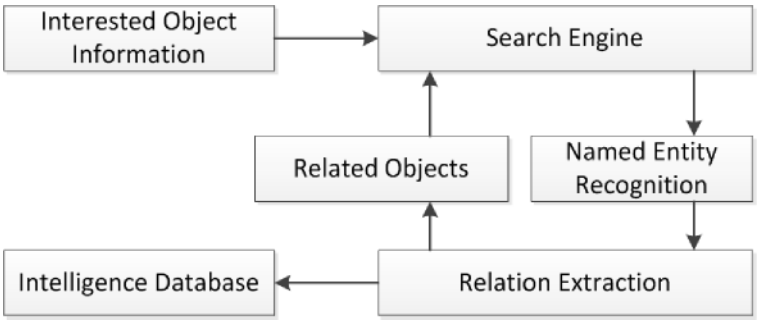

Fig. 2. The process of collecting multi-source intelligence from the Web

The multi-source intelligence is about a certain type of entity. It is collected from many Web sites and can construct a relatively complete view about the interested entity. For instance, we can collect information about “Oracle” from Wikipedia, BBS, the Oracle portal, and other sites to produce a systematic description about Oracle.

The system to collect multi-source intelligence consists of six parts, as shown in Fig.2. The major parts are the named entity recognition module and the relation extraction part. Within the named entity recognition module, it extracts named entities from Web pages on the basis of the rules predefined in terms of the interested objects. Tytpical named entities are person names, organization names, place names, merchant names, etc. The named entity recognition module is used to extract those special entities from Web pages. In the relation extraction part, it determines the relations among the named entities and therefore forms the conceptual view about the interested objects. The extracted entities as well as the relations among them are finally saved in a database, which can be used for further intelligence analysis.

2.3 Strategic Intelligence

The strategic intelligence is about the competitive strategy of a certain company or other organization. This type of intelligence is the most difficult one to obtain in the Web, as we have to conduct deeply analysis on the collected single-source and multi-source intelligence and usually will depend on some analytic models and tools. However, it is most valuable for governments or companies, as they can know the potential developing strategies of their competitors.

The strategic intelligence is based on the single-source intelligence and the multi-source intelligence. It is an integrated description about a set of objectives. For example, the strategic intelligence about Oracle may consist of the description about Oracle's profile, its business relationships, its main competitors, its main products, and its recent events. The strategic intelligence system is to integrate the different parts of intelligence and forms a systematic view for the interested objects. Fig.3 shows the basic process to generate strategic intelligence.

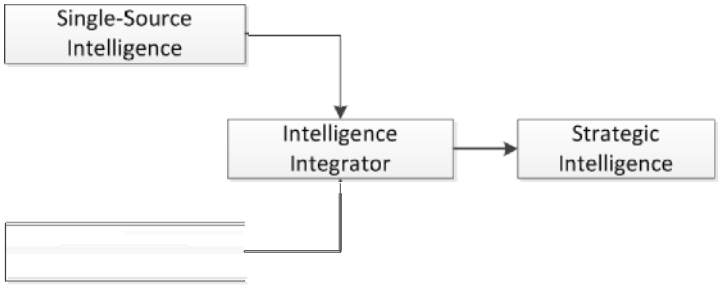

Fig. 3. The process of generating strategic intelligence

3 Design of a Real-Time Intelligence System

In this section, we discuss the design of a real-time system to collect intelligence from the Web.

3.1 Architecture

Fig.4 shows the architecture of a real-time intelligence collecting system. The system mainly contains five layers: data, data collection, single-source intelligence collection, multi-source intelligence collection, and strategic intelligence collection. Each layer consists of some specific algorithms to accomplish the task of the layer.

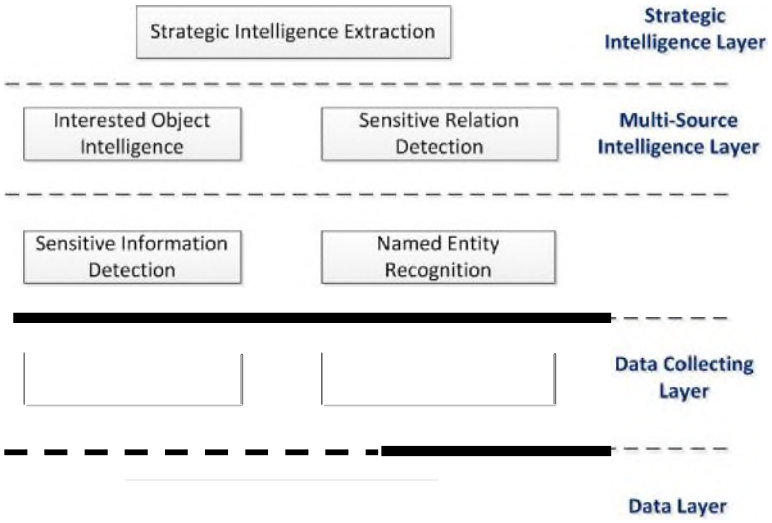

Fig. 4. The architecture of the real-time intelligence system

3.2 Key Technologies

3.2.1 Named Entity Recognition

Extracting entities from Web pages is one of the hottest issues in Web information extraction and retrieval [4]. The term is called Named Entity Recognition. Named entity recognition was first introduced as a subtask in the Message Understanding Conference (MUC) [5]. Its main task is to recognize and classify the specific names and meaningful numeric words from the given texts. Typical named entities are company names, person names, addresses, times, etc. Most of the previous research in this field focused on three types of named entities, namely time entities, number entities, and organization entities [6]. The major methods used in named entity extraction include rule-based approaches, statistical methods, as well as hybrid methods [4].

(1) Rule-Based Approaches

Those types of methods are the earliest ones used in named entity recognition. According to rule-based approaches, the rules of formulating named entities are first defined which are then used to match words in Web pages. There are some early sys-

tems built on this method, including NTU, FACILE, and OKI. The rule-based approach is usually associated with knowledge. For example, the NTU system uses knowledge about directional words and indicative words, and the FACILE system uses context knowledge, while OKI uses predicates knowledge in its implementation. The knowledge used in the rule-based approaches is very useful to improve the performance of named entity recognition.

(2) Statistics-Based Approaches

Like most artificial intelligence researches, the knowledge acquirement is a bottleneck in the rule-based approaches. Therefore, many people turn to use the statistical methods to perform the named entity recognition task. One of the main advantages of the statistics-based approaches is they can be directly used for different domains, in case that they are trained with the corpus in the new domain.

The typical models in the statistics-based approaches are the n-gram model, the HMM (Hidden Markov Model) model, the ME (Maximum Entropy model) model, the decision tree model, and other ones. The HMM model is widely used and has been demonstrated to have good performance.

(3) Hybrid Approaches

The rule-based approaches are very subjective and can not suit for different applications. On the other side, the statistics-base approaches usually introduce a very large searching space. So the hybrid method is proposed, which is a combination of the previous methods.

3.2.2 Business Relations Extraction

Business relations are very important for companies. Generally, there are several types of business relations. The ACE (Automatic Content Extraction) has defined six types of relations in English texts [7]. However, those relations are not defined for competitor intelligence. The only interested types in ACE are the Person-Social relation and ORGAffiliation relation. But these relations are too rough for business relations intelligence extraction.

Generally, the business relations can be classified into two types: Inner-ORG relations and Inter-ORG relations [8]. The Inner-ORG (ORG is the abbreviation of the word "organization") relations refer to the business relations between a company and its components, e.g. company manager, company-employee, and so on. The Inter-ORG relations are relations among different companies. Examples of the Inter-ORG relations are company investor, company-supplier, company-partner, etc.

(1) Inner-ORG relations

The Inner-ORG relations refer to the business relations among the entities of the same organization. A lot of information about a company can be extracted from the Web, e.g., name, address, email. This task is somehow easy to perform, because many methods have been proposed to extract different named entities [6]. Typical named entities are company names, person names, addresses, times, etc. Most of the previous research in this field focused on three types of named entities: time entities, number

entities, and organization entities. According to the context of competitor intelligence extraction, several types of named-entities are needed to be studied. However, we can use previous approaches to extract the named entities needed in the extraction of Inner-ORG relations. We can further classify the Inner-ORG relations into four types, which are ORG-person relations, ORG-location relations, ORG-time relations, and ORG-statistics relations.

(2) Inter-ORG relations

The Inter-ORG relations refer to the business relations between two companies. With the development of virtual enterprises and enterprise union, the relationships among different companies become more and more important in the market competition. Therefore, it is very important to recognize the competitors' business relations with other companies. Typical Inter-ORG relations are the relations among the companies who are contained in the same supply chain. For example, who are the suppliers of Lenovo? We can classify the Inter-ORG relation into four types of relations, which are cooperation relation, invest relation, sales relation, and supply relation.

4 Conclusion

With the rapid development of Internet and Web 2.0, there are a lot of intelligence which can be found in the Web. In this paper, we present an ontology about the intelligence in the Web, which consists of three types of intelligence: single-sources intelligence, multi-source intelligence, and strategic intelligence. We discuss the initial algorithm to collect those intelligences from the Web, and further present a system to acquire intelligence from the Web and via a real-time way. The key issues are also discussed in the paper.

Acknowledgement. This work is supported by the National Science Foundation of China under the grant no. 71273010 and the National Science Foundation of Anhui Province (no. 1208085MG117).

References

1. Khoury, I., El-Mawas, R., El-Rawas, O., et al., An Efficient Web Page Change Detection System Based on an Optimized Hungarian Algorithm. IEEE Transaction on Knowledge Data Engineering (TKDE), Vol.19(5), pp.99-613, 2007
2. LaMar, J., Competitive Intelligence Survey Report, In: http://joshlamar.com/documents/CIT Survey Report.pdf, 2007
3. Mikroyannidis, A., Theodoulidis, B., Persidis, A., PARMENIDES: Towards Business Intelligence Discovery from Web Data, In Proc. Of WI, pp.1057-1060, 2006
4. Khalid, M., Jijkoun, V., Rijke, M., The Impact of Named Entity Normalization on Information Retrieval for Question Answering, In Proc. of ECIR'08, pp.705-710, 2008

5. Sundheim, M., Named Entity Task Definition-Version 2.1, In Proc. Of MUC, pp.319-332, 1995
6. Whitelaw, C., Kehlenbeck, A., Petrovic, N., et al., Web-scale Named Entity Recognition, In Proc. of CIKM'08, pp.123-132, 2008
7. ACE (Automatic Content Extraction) English Annotation Guidelines for Relations, Version 6.2, (2008) Linguistic Data Consortium, In: http://www.ldc.upenn.edu/Projects/ACE/
8. Zhao J., et al., Business Relations in the Web: Semantics and a Case Study. Journal of Software, 5(8): 826-833, 2010

