http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (12): 2343-2349, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.2343.2349

Web Intelligence Analysis in the Semantic Web Based on Domain Ontology

[1]Chunnian Liu, [1]Dehui Yang, [1]Yonglong Wang and [2]Qin Pan
[1]Department of Information Management, NanChang University, China
[2]College of Economics Management, Huazhong Agricultural University, China

Abstract: Present study shows how the semantic web and ontology technologies are used in the web competitive intelligence analysis process. Most of the intelligence resources are on the Web, so it is very important to describe the semantic meaning of the web resources. In present study, we used the XML, RDF and Ontology to describe the intelligence resources. Besides, building a competitive intelligence domain ontology is the precondition of the intelligence analysis. On the basis of the domain ontology, a frame of the semantic competitive intelligence analysis system is proposed which consists of five main components. According to the system, the intelligence analysts and other users can directly access to the competitive intelligence database and retrieve the resources they need by concepts. The study conclude the description of the implementation strategies of the system model.

Key words: Domain ontology, semantic web, intelligence analysis, web resources

INTRODUCTION

Business competitive intelligence can be defined as the process of monitoring a firm's external environment to obtain information relevant to its decision-making process. In today's society, the competitive intelligence resources have not been limited to traditional paper books, journals, minutes and patents. Web information and a variety of digital resources have become one of the most important parts of the intelligence resources. Internet has become a large repository of information that could be relevant to a firm's decision making. In late 2004, Google announced that they had indexed more than 8 billion Web pages. For instance, looking into the Web sites of a firm's competitors can reveal useful information about the firm's competitive environment (Hsinchun *et al*., 2002). Therefore, web intelligence analysis becomes a necessary part of the intelligence work. Generally, the Web analysis relies on three general sets of information given: past usage patterns, degree of shared content and inter-memory associative link structures. The Web sites of other stakeholders of the firm, like customers, suppliers and pressure groups, often have hyperlinks pointing to each other or are pointed to by the same set of Web sites, they are known as the "Web communities" of the firm. The identification of Web communities is important in the web intelligence analysis process and the common tools to search and analyse the information of "Web communities" are mainly on the basis of traditional keyword-based search on the Web. However, keyword-based searching can only return search results with a page containing the search keyword but does not guarantee the relationship between the search result and the firm of interest. Hence, in the face of geometrically increasing web resources, how to get valuable intelligence from the massive information and obtain the favorable results of decision-making, has become an urgent problem for the intelligence analysts and even more people.

In recent years, many tools have been developed to incorporate more than one of the functions of searching, analysis and visualization. For example, Wang *et al*. (2005) propose a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. A tool called "Redips" is presented to automatically integrate backlink meta-searching and text-mining techniques, in order to facilitate users in performing such business intelligence analysis on the Web (Chau *et al*., 2007). Besides, an EBizPort system was designed to address information needs for the business/IT community and the collection-building process of EBizPort is designed to acquire credible timely and relevant information (Marshall *et al*., 2004). It can clearly be seen that current web intelligence analysis tools are often good for information retrieval but lack the functionality to find the Web communities of a firm, especially in the increasingly complex network environment. So then a variety of IT technologies, such as the semantic web technology, are introduced and widely used to collect and analyse the information resources.

Corresponding Author: Chunnian Liu, Department of Information Management, NanChang University, China

Present study, we apply the Semantic Web technology to analyse the web business intelligence which aim to facilitate access into the web resources for the intelligence personnel. We construct an ontology-based semantic intelligence analysis system, on the basis of analysing the semantic description of web resources and building a competitive intelligence domain ontology.

SEMANTIC DESCRIPTIONS OF WEB INTELLIGENCE RESOURCES IN THE SEMANTIC WEB

Overview: The web intelligence resources from public or semi-public sources include the following data formats: database, text, images and multimedia and so on. According to complexity, these intelligence resources can be divided into three levels. Level I is the most complex data format and its expressing content is the most abundant, including audio, video, pictures and other documents but difficultly understood by the computers. Level II includes all kinds of text in natural language to express the data content. The data in level III can be described by the database fields and easily understood by the computers.

In the process of the web intelligence analysis, we firstly need to transform the data from a simple format to a complex format gradually, in order to achieve a unified data format of information resources, which is conducive to the further semantic description of the web resources. The data resources in the multimedia and image format are firstly converted to text data by using artificial way, voice recognition technology, image extraction technology and so on. Then, we mainly use information extraction technology, to achieve the format conversion from the text to the database format. When the data resources of the level I and level II are converted to level III in the database fields, there will be a lot of data in the level III, involving a large number of entity definitions and the relationship definitions between entities. Thus, the Semantic Web technology is one of the main technologies to achieve the semantic description of resources.

Semantic Web is a kind of diagram representation based on the web knowledge which can implement the semantic structure of knowledge on the Web and describe the related knowledge among the concepts and rules (Poibeau and Dutoit, 2002). The aim of the semantic web is to make the information on the web understood by the computers and then facilitate access to the web resources for the intelligence personnel. TimBerners-Lee (Yong-Chao and Jun-Min, 2007) first proposed the concept of semantic web in 2000, meanwhile raised the level structure of the semantic web, respectively from the bottom to the top: UNICODE and URI, XML, RDF, Ontology, Logic, Proof and Trust.

Semantic descriptions of web Intelligence resources: From the perspective of the hierarchy structure of Semantic Web, XML, RDF and Ontology are the main core of the Semantic Web. At the same, they are also the key of semantic description of Web intelligence resources.

XML: XML, as a language of resource description has good flexibility and scalability which is considered as the data exchange standards in the future Web. In addition to add a description tags to the data, XML is also the important supplement to HTML. Not only can XML provide the expression of the content of the resources but also can it provide the structural information that the resources need. When describing the semantics, the data semantics of XML is mainly hidden in the structure and tags. So we can express the data semantics and relationships between data according to concatenation and nestification among data. For example:

```
<search Engine>
<name> Google </ name>
<URL> www.google.com </ URL>
</ search engine>
```

It can be seen that the data expression of XML is a tree structure in the semantic description of resources. This kind of tree structure lack of some flexibility in the description of metadata and also may cause the semantics missing to the description of Web intelligence resources in network structure. Besides, as XML pages contain a large number of information in the format of image, multimedia and text etc., they can not be directly handled by intelligent software agents. So the metadata is needed to describe the XML resources. For that the Semantic Web must provide a more simple and effective resource description framework or model to describe the metadata.

RDF: RDF which is a draft of describing and dealing with the metadata, is a framework of resources description by XML tags. It is independent of any language and field. RDF which is the basis of dealing with the metadata, can provide the information that the machine can understand for the interaction between application programs on the Web. RDF express the Web intelligence resources as a graph in the form of XML markup by using the triples (i.e., subject, property, object). According to URI (Uniform Resource Identifiers), RDF can point the specific location of the document or object on behalf of Web information

resources for the computer. In addition, RDF mainly describe the Web information resources by the properties and values of properties. The biggest difference between RDF and XML is the expression of semantics. In other words, XML is an extensible markup language and RDF is a knowledge representation language on the Web which is a special form of predicate logic. Both of XML and RDF are the core of the Semantic Web and can provide semantic description for the Web intelligence resources. But neither of XML and RDF can solve the problem of polysemy.

Ontology: In order to solve the problem of polysemy in the semantic description of XML and RDF, we need to introduce ontology. Ontology determine the precise semantics of the concept by a strict definition of the concept and the relationship between concepts, to express common recognized and shared knowledge (Zhihong *et al*., 2002). Ontology is the key factor to achieve the semantic interoperability in the Semantic Web and also the foundation of solving the information sharing and exchange on the semantic level. The specific semantic expression of ontology needs the description languages. Currently, there are a variety of ontology description languages, such as OWL, ontolingua, Loom and so on. Among them, OWL is the recommended standards of W3C and can support interoperability of XML, RDF resources between machines. Compared withe XML and RDF, OWL can better describe the semantics of Web information resources. Meanwhile, OWL can also describe the relationships between the intelligence concepts, such as the inclusion relationship, the relationship between individual and general. Therefore, ontology plays an important role in the process of describing the semantics of Web intelligence resources in the context of the Semantic Web. So next, we will build a domain ontology of competitive intelligence which could be very beneficial to the work of Web intelligence analysis.

In a word, the Semantic Web uses XML to define the format of label and the flexibility of RDF to express the data and then describes the clear meaning of terms and their relationships in the network documentation by using a web ontology language (such as OWL).

CONSTRUCTION PROCESS OF THE COMPETITIVE INTELLIGENCE DOMAIN ONTOLOGY

Overview of the domain ontology: Competitive intelligence analysis is conducted under specific purpose, scope and needs. So intelligence analysis can not just be direct processing of intelligence resources but analysis of Web information resources in the specific background and associated knowledge. Present study presents a competitive intelligence domain ontology, to model the issue background and associated knowledge of competitive intelligence analysis.

The Competitive Intelligence Domain Ontology (CIDO) is defined as a tuple:

$$CIDO = \{C, R\}$$

where, C represents the concept set in the field of competitive intelligence and R represents the relationship set between the concepts.

In the competitive intelligence domain ontology, Concept can be expressed as a quintuple, including Name, Label, RelationSet, SynonymsList, Description. Relationship in the domain ontology mainly consists of the following four types of relationships (Liang, 2008): kind-of, instance-of, part-of and attribute-of. Among the relationships, kind-of relation is the basic relationship between concepts. For example, the product profile is a kind of competitor information. Instance-of relation represents the relationship between the instances and concepts. For instance, Kangshifu enterprise is an instance of the Tongyi's competitors. Part-of relation informs the relation between part and whole, such as the competitors' intelligence and the competitive intelligence. At last, attribute-of relationship tells that one concept is an attribute of another concept. For example, name is an attribute of a product. Base on the domain ontology, we can not only describe the resources of the specific area in the semantic level but also make the intelligence resources in the area increase from the content level up to the semantic level, in order to make the analysis and management of the information resources more effective and intelligent.

Construction of the competitive intelligence domain ontology: By summarizing the experiences of enterprise competitive information analysis, we expect to carry on the abstract summary of the enterprises' competitive information with the aid of the domain ontology to organize a higher level knowledge abstracting. The competitive information analysis establishment is on the foundation of the domain ontology which guarantees the knowledge exchange uniformity and subsequently realizes the intellectualized web competitive information analysis through the associated heterogeneous modules. Constructing the domain ontology of competitive intelligence is the key aspect of semantic intelligence analysis. Gruber proposed five principles of ontology construction in 1995 as the follows: clarity and objectivity,

integrity, consistency, maximum one-way scalability, the least constraint. These principles have been widely applied to the ontology construction of various areas. Based on the above principles, the methods of ontology construction include TOVE, METHONTOLOGY, skeleton, KACTUS, SENSUS, IDEF5, seven-step method and so on. As the skeleton method is designed to build enterprise ontology, so this paper chooses the skeleton method as the construction method of the competitive intelligence ontology. In addition, the building tools of the domain ontology mainly consists of ontolingua, WebOnto, OntoSaurus, protégé, OilEd, OntoEdit, etc. Each of the above construction tools has its own advantages and each of them is irreplaceable, while in this present study we use protégé to conduct the competitive intelligence domain ontology.

On the basis of the enterprise ontology, we use the skeleton method to build the competitive intelligence domain ontology which is a set of commercial terms and definitions among enterprises. The process of skeleton method (Abburu and Anandhi, 2010) includes the following several steps, shown in Fig. 1. First, determine its corresponding domain scope. The "competitive information" is the highest point as we plans to discuss the domain ontology. It consists of the "enterprise competitive information", the "military competitive information", "the market competitive information"and so on. Next, define all the terminologies and their relations in the domain ontology with the domain experts' help, namely the "enterprise competitive information" belongs to the "competitive information" with the attributes "the competitor", "the environment of competition", "the competition strategy" and so on. An explicit and formalized standard explanation should be formed through constructing a group of glossaries, definitions, axioms, theorems and so on associated with entity sets. Then ontologies are indicated based on the segment semantic model. Finally, ontology evaluation criteria are established and an appraisal is carried on to this domain ontology.

In order to more clearly explain the construction of the proposed competitive intelligence domain ontology, an instance of competitive intelligence domain ontology, called the "sales strategy of Tongyi100 instant noodle", is shown in Fig. 2.

Web intelligence analysis system based on the domain ontology: Competitive Intelligence system can assist business executives to make decisions and business development strategies. In a dynamic competitive environment, its role is increasingly important. Competitive intelligence analysis system is a core subsystem of the enterprise competitive intelligence system. Its quality of design and development will directly affect the strategic decisions of senior managers.

Frame and components of the system model: The aim of present study is to develop a web intelligence analysis system based on the domain ontology. The system, shown in Fig. 3, consists of several components that together fulfil the required functionality, including domain knowledge modeling, ontology importing and mapping, spidering and tagging, analysis of semantic content, retrieval of competitive Intelligence. In the following, each of the above components is introduced.

Domain knowledge modeling: Ontology is an important tool of knowledge modeling which has been widely used in text processing, semantic web and other fields. As a knowledge modeling tool, ontology can represent the network resources by directed graph with labels and it is also suitable for logical reasoning. At present, the intelligence resources on the Web is increasing in the speed of geometric progression which causes that it is very difficult for intelligence analysts or other users to access intelligence information they need. Therefore, using ontology as the modeling tool of competitive intelligence semantic content, is an important development direction of the competitive intelligence analysis area. The competitive Intelligence domain ontology that present study proposed can provide necessary support base for the semantic content analysis and retrieval of competitive intelligence information.

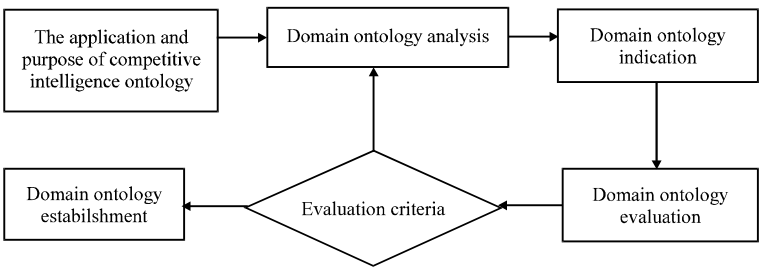

Fig. 1: Construction process of the Competitive Intelligence Domain Ontology

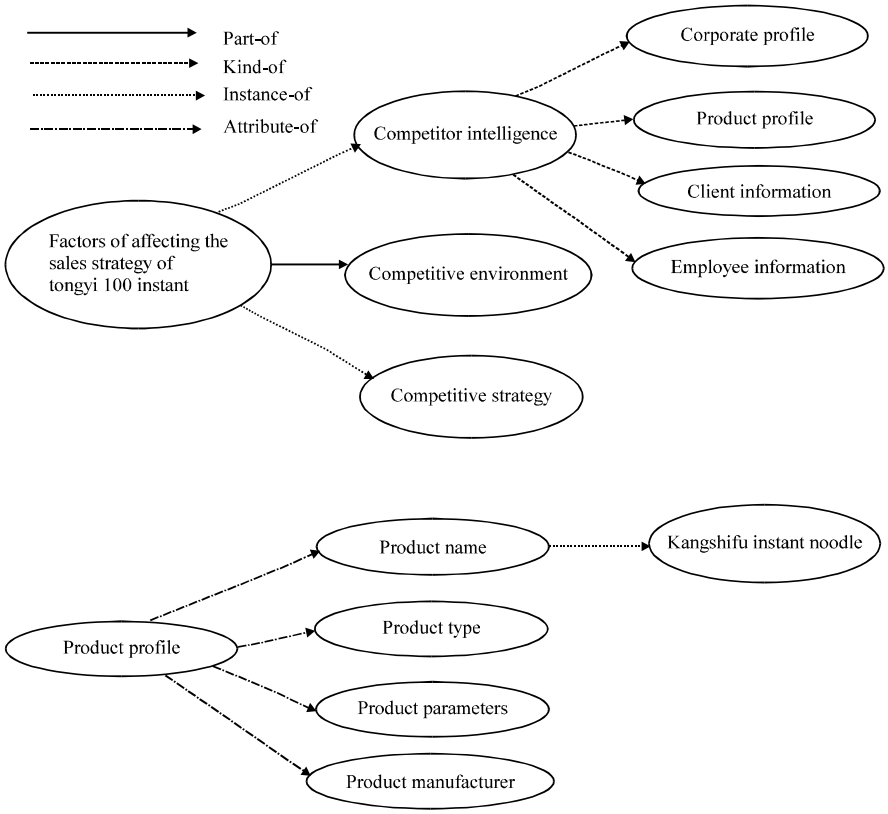

Fig. 2: The "sales strategy of Tongyi100 instant noodle" domain ontology

Ontology importing and mapping: The domain ontology constructed above have to be imported to the system by using an RDF-parser. RDF is a data format based on XML which is well suited as an interchange format of ontologies and controlled vocabularies. So RDF is widely accepted for the ontology importing. Importing an ontology mainly consists of three steps. First, convert the RDF-file to JENA-model. Second, read the configuration file which stores individual properties of an ontology and maps the different notations for equivalent characteristics, for importing ontologies. Third, write the concepts, relations and synonyms in the ontology into a database.

Mapping ontologies means that the equivalent concepts of the different ontologies are aligned. The purpose of this ontology alignment procedure is that only equivalent concepts should be aligned, even if not all the equivalent concepts could be found. For this aim, the approach of "comparing the concept context between the ontologies to be aligned" seems to be best suited.

Spidering and tagging: To download the intelligence resources for the Web and mirror them in the system, the webspider is a suitable tool. The spidering process involves several steps: removing the special characters, droping stopwords and badwords, calculating tag, droping HTML tags, writing words into the intelligence resource database. After that, we have to tag the intelligence resource in the database according to the domain ontology.

Analysis of semantic content: Semantic content analysis of competitive intelligence, as a core component of the competitive intelligence analysis system, is to obtain the high-level semantic of competitive intelligence assisted in the domain ontology, namely concepts of competitive intelligence. Semantic content analysis of competitive intelligence includes two levels: the low-level semantic extraction and the high-level semantic concept detection. Low-level semantic extraction mainly consists of syntactic segmentation, matching, classification and detection of perceptual concepts. High-level semantic concept detection includes metanotion detection and high-level concept detection.

Retrieval of competitive Intelligence: Retrieval of competitive Intelligence means that intelligence analysts

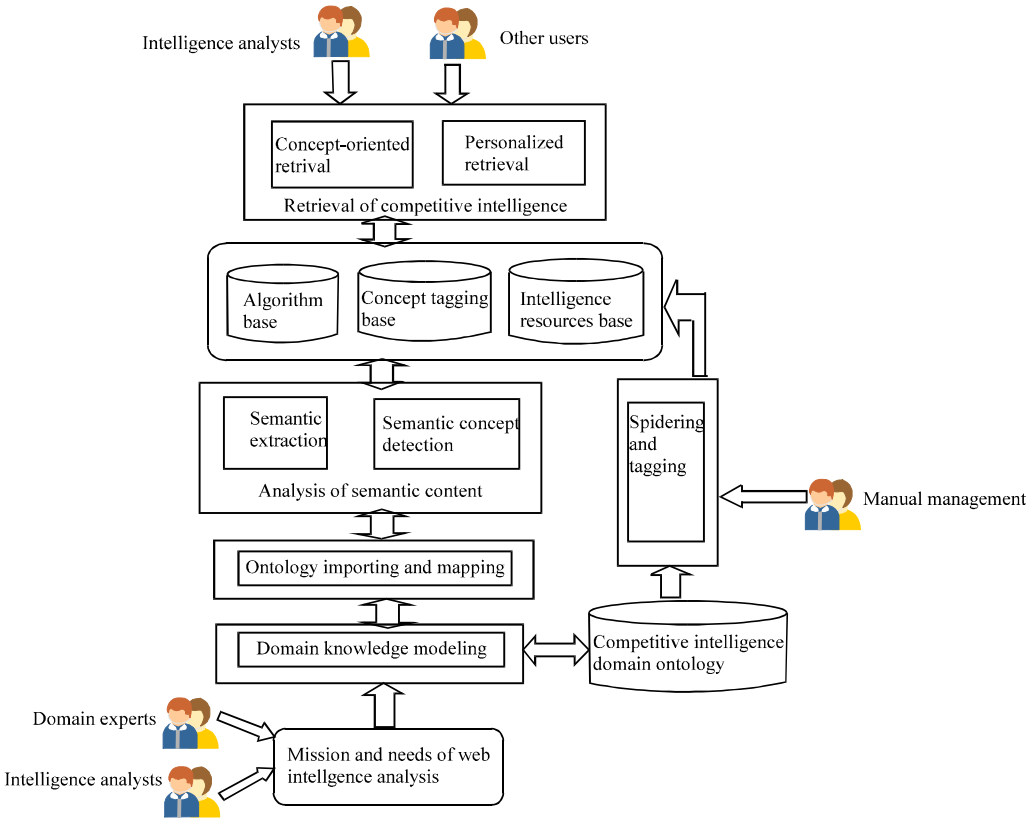

Fig. 3: Model frame of domain ontology-based web intelligence analysis system

and others apply the methods of intelligence analysis to obtain information knowledge on the basis of the domain ontology and the semantic content analysis. The main analytical techniques of competitive intelligence include statistical analysis, association, classification, clustering and so on. Competitive intelligence retrieval provides the users with concept-oriented and personalized intelligence database access methods. On this basis, intelligence analysts can browse and search the competitive intelligence resources according to mission requirements or personal preferences.

Prototype implementation of the system: Since the domain ontology is the results between the concepts and support logical reasoning, we take the domain ontology as the core of the web intelligence analysis system which can realize the knowledge exchange uniformity between isomerism modules on the foundation of the standard semantics. Meanwhile, we construct a kind of intellectualized user's model on the basis of the ontology, to make web intelligence mining better meet users' personalized demands. In the following, present study will explain more clearly the prototype implementation of the web intelligence analysis system.

First of all, intelligence analysts need to clear the mission and needs of competitive intelligence analysis and build the domain ontology with the help of domain experts. Then, the domain ontologies are imported to the system and the equivalent concepts in different ontologies are aligned to eliminate the problem of polysemy. Again, webspider which is a spider tool, is introduced to collect web competitive intelligence resources. Based on the ontologies, artificial marking is done to create a competitive intelligence resource base. The next part is the core phase of the competitive intelligence analysis, namely semantic content analysis of competitive intelligence. At that stage, the intelligence content of different analysis needs, will produce corresponding training data and training algorithm, etc. In accordance with the relevant training contents, analysts extract the perception concepts and detect the high-level concepts, to obtain the results of semantic content analysis. If there is a big gap between the results of semantic content analysis and the users' requirements, then it is necessary to return to the training phase and repeat the next operations, until the intelligence analysts or other uses are satisfied. Finally, in the stage of competitive intelligence retrieval, according to the query

needs and task requirements, intelligence analysts and other users can easily access the competitive intelligence database which has been manually tagged by the intelligence concepts and the system will return certain competitive intelligence the users needed.

CONCLUSIONS

Because the domain ontology has a good concept hierarchical structure and supports logical reasoning, it provides certain references for realizing the competitive intelligence analysis. In this publication, a domain ontology of competitive intelligence is built, in order to support the process of semantic content analysis and the intelligence retrieving on the Web. Based on that, a semantic competitive intelligence analysis system is presented which involves the following components, such as the domain ontology, semantic content analysis and intelligence retrieving and so on. The methodology presented in this study should be revised in the future when the competitive intelligence analyzing system is better developed and related information is more readily available. Further research is also needed on the possible benefits of establishing enterprise competitive intelligence mining system and the contexts that are suitable for specific districts to promote the level of intelligence analysing.

ACKNOWLEDGMENTS

Science and Technology project of Education Department in Jiangxi Province "Ontology-based Agricultural Disaster Emergency Information Integration Services: Theory, Methods and Applications" (GJJ11273); Jiangxi Arts and Sciences planning issue, "Research on Cultural Information Service Alliance of Public Libraries in Poyang Lake Ecological Economic Zone" (YG2010028).

REFERENCES

Abburu, S. and R.J. Anandhi, 2010. Concept ontology construction for sports video. Proceedings of the 1st Amrita ACM-W Celebration of Women in Computing in India, Sept. 16-17, ACM, New York, pp: 425-425.

Chau, M., B. Shiu, I. Chan and H. Chen, 2007. Redips: Backlink search and analysis on the Web for business intelligence analysis. J. Am. Soc. Inform. Sci. Technol., 58: 351-365.

Hsinchun, C., M. Chau and D.D. Zeng, 2002. CI Spider: A tool for competitive intelligence on the web. Decision Support Syst., 34: 1-17.

Liang, B., 2008. Research on video intelligence analysis using ontology. Graduate School of National University of Defense Technology, Changsha, China.

Marshall, B., D. McDonald, H. Chen and W. Chung, 2004. EBizPort: Collecting and analyzing business intelligence information. J. Am. Soc. Inform. Sci. Technol., 55: 873-891.

Poibeau, T. and D. Dutoit, 2002. Generating extraction patterns from a large semantic network and an untagged corpus. In proceedings of COLING-02 on SEMANET: Building and using semantic Networks, pp: 1-7.

Wang, X., A. Abraham and K.A. Smith, 2005. Intelligent web traffic mining and analysis. J. Network Comput. Appl., 28: 147-165.

Yong-Chao, L. and L. Jun-Min, 2007. Reasoning on ontology in semantic Web. Comput. Technol. Dev., 17: 101-103, 107.

Zhihong, D., T. Shiwei, Z. Ming, Y. Dongqing and C. Jie, 2002. Overview of ontology. Acta Scientiarum Naturalium Universitatis Pekinensis, 38: 730-738.