Importance of User in Extracting Knowledge from Web for Ontology Construction

A De S. Jayatilaka, Dr. G.D.S.P .Wimalaratne, University of Colombo School of Computing

Key Words : *Semantic Web, Web Mining, Ontology, Concepts, Relationships*

The introduction of the World Wide Web has given rise to an enormous amount of information to be accessed in the digital form. However, exponential growth of these digital documents have given rise to many new challenges, specially in terms of information retrieval. These traditional web pages are not descriptive enough to express themselves and are over loaded with information. Therefore, web search results suffer from problems of accuracy and precision. In addressing this issue, incorporating semantics to these documents was identified as the key element.

All present research in terms of extracting ontologies only focus on web content. However, through this paper we argue that web usage patterns also need to be taken in to account in order to improve the accuracy of the ontology. Web user's perspective and the web author's perspective on the same set of web pages can be different. Web authors explicitly use hyper-links to link web pages which are conceptually related. However, web users might observe other conceptually related web pages. These are reflected implicitly in the user navigation patterns. The main goal of this research is to use this hidden knowledge in the semi automatic ontology construction process.

The comprehensive qualitative and quantitative evaluation carried out has shown positive results. Therefore, this solution is intended to be usable for transformation of large web document corpus to semantic web.

1. Introduction

Semantic web is considered as the second generation of the World Wide Web (WWW). Tim Berns Lee defines semantic web as "web of data that can be processed directly or indirectly by machines". Semantic web is a web with a meaning. The information on semantic web is readable to both humans and machines.

Ontologies are considered as the pillars of semantic web. Ontology is a term in philosophy that means theory of existence. Ontologies with respect to computer science are the formal representatives of set of concepts and relationships between those concepts within a domain. Learning ontologies from web documents is a very challenging task. One option is to manually handcraft the ontologies. However, this is not a practical solution because manual development of ontologies is a very time consuming, tedious and a costly task. Many resources such as text, dictionaries, thesauri, databases are used in the ontology learning process.

Most of the time ontologies are handcrafted. However, manual construction of an ontology requires a significant amount of domain knowledge. Therefore it a very costly task in terms of time and money [1–3]. Therefore, it has become a bottleneck for the fast growth of semantic web. As a result, automating the process of ontology construction is highly important in order to move forward in this digital era.

Most of the research up to date only considers web content for ontology development. However, in this way we only take in to account the web designer's view. However, these approaches fail to identify the semantics from the web user's perspective. Web user's might view other conceptual relationships which are not reflected in the web content. Therefore, this research highlights the importance of taking in to account this hidden knowledge in ontology development process.

In this context, we try to propose a novel approach to extract concepts and their relationships from semi- structured web pages by using a combination of web mining techniques namely, web content mining and web usage mining. These extracted conceptual relationships form a semantic network which could be directly used to develop the semantic web.

2. Related Work

Manual extraction of conceptual relations and manual construction of ontologies is a time consuming, tedious and a costly task. Therefore, conceptual relationship extraction should be made automatic up to a certain extent. However, full automation of the conceptual relationship extraction process will make the results inaccurate. Therefore, semi automatic approaches are mostly suited [1], [4].

There have been several research attempts to meet the challenge of automation of the conceptual relationship extraction process for web documents. The research attempts that focus on un-structured web pages [5], [6] with free text mostly use natural language processing techniques and simple text mining in the ontology learning process. There are only a few references to research attempts that focus on semi structured web pages [7], [8].

The importance of using web mining techniques when moving towards semantic web is illustrated in [1], [3]. However, most of these research attempts lack the focus on the user perspective in ontology development and only focus on the web content. Furthermore, [1], [3] explain that it is expected in the future to combine several web mining techniques together in the process of extracting semantics from web data and discuss the pathway to transform from the traditional World Wide Web (WWW) to the semantic web and discuss the role of web mining techniques in facilitating ontology development.

3. Web Author's View Vs Web User's view

Web authors and users could conceptually view the web site/web pages in different ways. The web authors define a navigation structure according to how they view the web site. However, it cannot be assured that web users go through the same procedure as suggested by the web designers. Sometimes, users move from one web page to another even without having direct hyper links among them [3], [9], [10]. The main reason for such movement is the users identify conceptual relationships that exist among those web pages. Therefore, an important design concern is incorporating web designer's view and web user's view in the conceptual relationships extraction process.

4. Architecture

Overall goal is to extract conceptual relationships from semi structured web documents. The *figure 1* illustrates the system architecture. The proposed methodology combines two web mining techniques namely, web content mining and web usage mining in the process of extracting conceptual relationships extraction process.

The project attempts to utilize the semi structured nature of the web pages using web content mining. A web site is designed by a web author and web developer. Web content mining is used to extract conceptual relationships that reside inside web documents. These conceptual relationships represent the web author's/web developers' ideas and thoughts and how they conceptually view the web pages. However, the users see the web site in a different perspective. Content mining cannot be used to capture these conceptual relationships that reside in the users' thoughts when he is navigating. Web content mining is only capable of extracting conceptual relationships according to the web author's view.

Web users navigate through the web pages according to their preference and according to how they view the web site conceptually. The web user navigation patterns can be identified using web usage mining and this knowledge is then used in the ontology development.

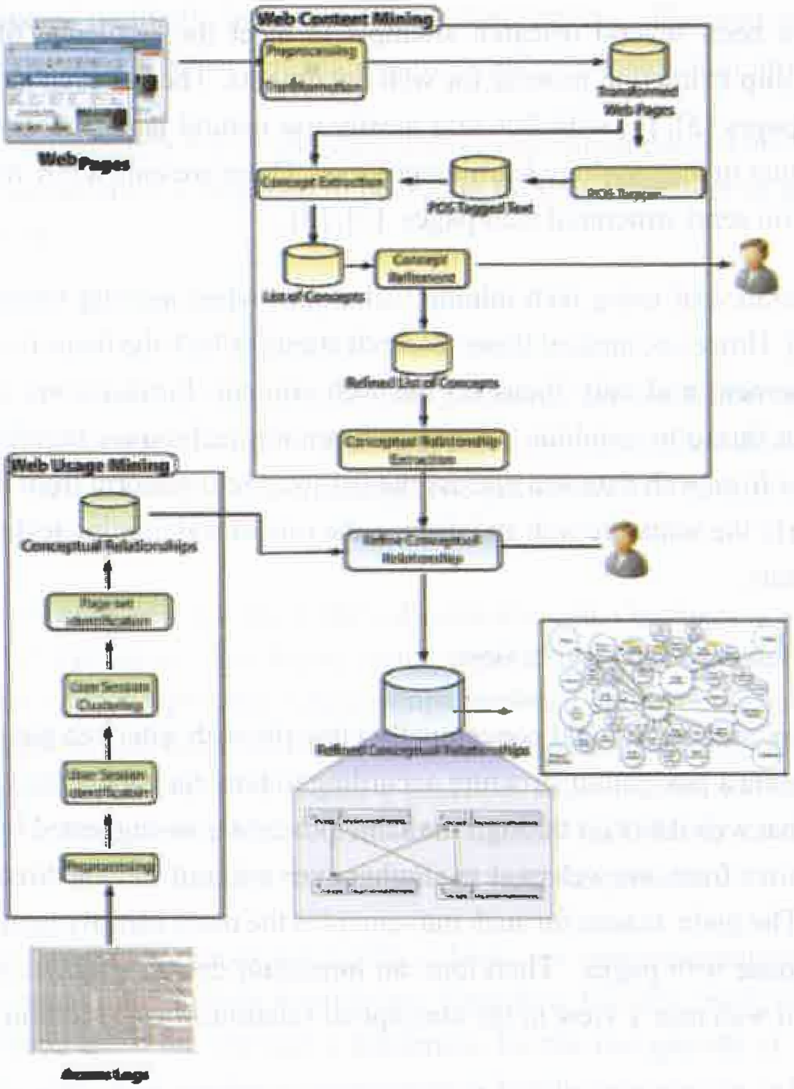

Figure 1 : Design Methodology

4.1 Incorporating web content in semi automatic ontology construction

Web content mining starts by pre-processing the web pages. Most of the web pages do not confirm to valid HTML syntax. The ill formed HTML tags should be pre-processed prior to page parsing. Then the web pages are parsed and important data is extracted and transformed in to a logical structure. Extracting conceptual relationships from web content could be categorized in to two phases namely, concept extraction, conceptual relationships extraction. This is discussed in detailed in [11].

4.2 Incorporating web usage patterns in semi automatic ontology construction

Web usage mining is the process of finding out what the users are looking for on web data [12], [13]. This information reveals information about the users search patterns for particular group of people belonging to a particular group. Web usage mining information is widely used in e-commerce application to increase sales through personalizing the web site layout and the structure based on the history of the access sequences. However, it is not researched much how to improve the semi automatically developed ontologies through web usage mining.

This project uses web usage mining to extract conceptual relationships that could be learnt about the web site/pages according to the usage patterns. These extracted conceptual relationships are used in the conceptual relationship refinement stage along with the conceptual relationships extracted by mining the web content in the web site/pages.

However, relying only on the information generated by web usage mining is not a practical because it only reveals the conceptual relationships that could be extracted from web pages according the users perspective. Furthermore, if the web site has a rapid changing nature where the content is changed frequently identifying user navigation patterns may not reveal adequate information. Therefore, web usage mining is used only to refine the conceptual relationships extracted by mining the content of the web pages.

4.3 Preprocessing access logs

Access log stores all the records of requests for individual files by the people [13]. Well designed data mining techniques could obtain important information through the access logs. These logs in general contain the IP address of the requester, the user name of the user who generated the request (If applicable), the date and the time of the request, the method of the request(GET or POST), the results of the requests, the size of the data sent back. A log entry is automatically added to the log file each time a request has been made. It is important to understand that entries of all users are mixed up in the log file and access logs contain raw data. These raw data in access logs have to be cleaned before extracting knowledge because not all data is important in decision making.

When the access logs are cleaned they could be used to provide useful information such as the number of visitors per day, how many requested each page in the web site etc. In this project preprocessing the access logs also includes removing duplicate entries. This is done using a duplicate time out value. The duplicate time out value species the maximum time in seconds between successive requests of the same document from the same host. Depending on the caching algorithm, the graphics might be requested multiple times in successive requests during a very short time. For our purposes, we only want to recognize multiple requests of the main document. It is also important to notice that we only need information about the main documents. Therefore, by having a filter path value we are dropping requesting to irrelevant objects such as images, scripts etc.

4.5 User session identification

The user session identification is a challenging task. The number of user sessions could determine the amount of traffic in the web site. Users could be identified based on the IP address. However, there are two main limitations in identifying user sessions based on the users IP address [14].

- One user could have several IP addresses even in the same session
- Several users could have the same IP address due to the effect of network address translation.

There are several methods that could be used to distinguish user sessions when several users have the same IP address. By using information stored in the \referee" and \browser" the different users having the same IP address could be distinguished. However, complete distinction is not possible. Cookies are also used for better user authentication. The default thirty minute session time out value is also used to break user click streams in to sessions. This value is based on studies done by [13]. Accurate user identification could be done using the user names; however, requiring users to authenticate is in appropriate in web browsing in general. Therefore, user names are not used in user session identification [14].

4.6 Clustering

Clustering is a technique to group together set of items that have similar characteristics. Clustering of user sessions will establish groups of users having similar browsing patterns. Such knowledge is heavily used for web personalization. In this research we cluster the users according to their navigation to extract information to facilitate the ontology learning process. We identified each user session extracted as a web transaction and we mapped these web transactions in to a multi dimensional space as a vector of URL references. K-means clustering algorithm was used to partition this space in to set of clusters based on the Euclidean distance function. Therefore, each cluster represents group of web transactions that are similar based on the co-occurrence patterns of the URLs.

In order to map the user sessions in to a vector space model we represent each web page visited as 1 and each web page not visited as 0 in the vector for a given user session. Some other research attempts suggests that rather than using binary weights in the vector space model it is better to use feature weights such as frequency of occurrence, the time spent on each web page etc [15]. However, [12] argue that feature based weights are not justifiable over binary weights when web transactions are considered. Furthermore [16] shows that the amount of time spent on a page generally is not a good measure of the interest of the user.

4.7 Association rule discovery

Association rule discovery finds groups of elements which are occurring frequently together in many transactions. Those groups are refereed to as frequent item sets. In this research we discover association rules based on the usage clusters. Apriori algorithm is used for association rule discovery. A usage cluster contains web transactions that have similar navigation patterns. Therefore association rule discovery will result in frequent item set representing each cluster. These extracted association rules relates web pages that are more often referenced together in a single user session. Therefore, association rules indicates sets of web pages that are accessed together with a support and confidence value that is exceeding some specified value. These web pages may not be directly linked with each other through direct hyper links. However, these page sets hint the existence of conceptual relationships.

4.8 Conceptual relationship extraction

Conceptual relationships are identified as suggestions using the extracted page sets. These suggestions are presented to the ontology developer and if he/she accepts these suggestions they are reflected in the semantic network. Conceptual relationships need to be extracted from the extracted page sets. However, there is not adequate information about the possible conceptual relationships only by considering the page sets extracted from web. usage mining. Therefore, Candidate conceptual relationships for the suggestions are identified by lowering the thresh hold value that was set for the relationship value at the web content mining. Then it is examined whether these candidate conceptual relationships are reflected in the page sets from web usage mining. If so, they are presented to the ontology developer as suggestions

4.9 Conceptual relationships refinement

The conceptual relationships refinement is done with the involvement of the ontology developer. The conceptual relationships extracted from mining the content of the web pages and the conceptual relationships extracted by mining the web usage patterns are presented to the ontology developer. The ontology developer can determine whether a particular conceptual relationship should be reflected in the semantic network or not. Furthermore, the ontology developer could add any new conceptual relationship to the semantic network.

5. Evaluation

In this research we have used two types of ontology evaluation techniques namely, 1) evaluation of the ontology against a gold standard ontology and 2) User evaluation.

5.1 Evaluation of the ontology against a gold standard ontology

We have used BBC wild life ontology as the gold standard to evaluate the semi automatically developed ontologies. As explained earlier there are no standard measures to evaluate or compare an ontology against another ontology. We have used two straight forward measures that are derived from Information retrieval namely, precise and recall. Precision is a measure of exactness where quantity of results returned is measured and the recall is a measure of completeness where quality of the results obtained is measured. When precision adopted to ontology learning tasks could be defined as the number of relevant concepts/conceptual relationships retrieved by the system dividend by the total number of concepts/conceptual relationships retrieved by the system. The recall measure when adapted to ontology learning could be defined as the total number relevant of concept/conceptual relationships retrieved by the system divided by the total number of the existing concepts/conceptual relationships that should have been retrieved. The positive feedback obtained using precision and recall values gave us some hints about how to gauge the thresholds for the system.

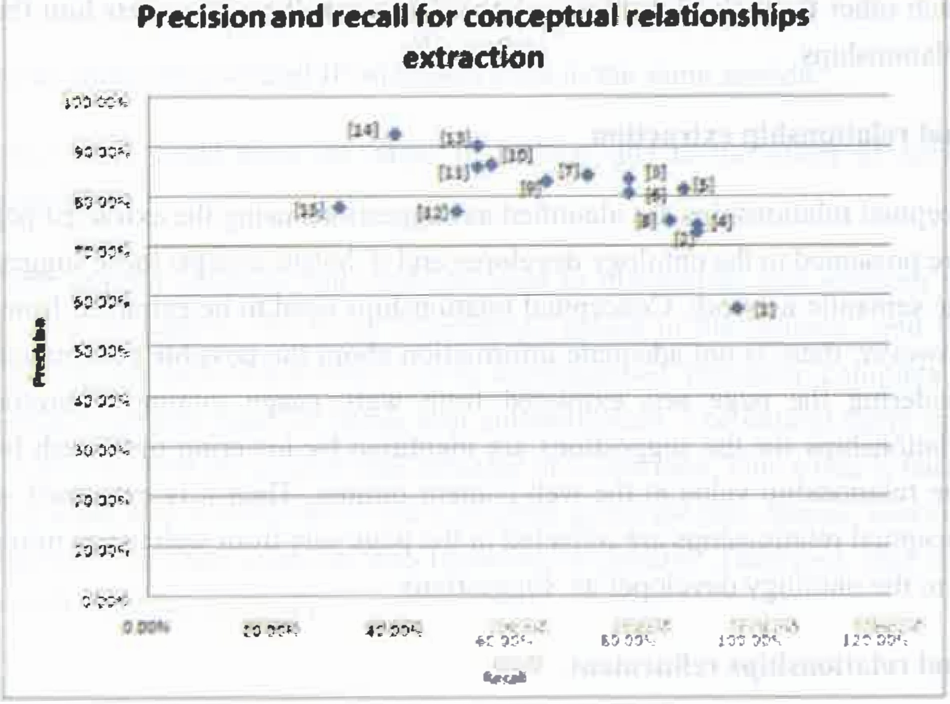

Figure : Evaluation Results

5.2 User Evaluation

Given the nature of this research, the user is the person who finally decides whether to include a particular concept/ relationship in the ontology. Hence he/she can assess the quality of the extracted concepts and the conceptual relationship Users were clustered in to three categories namely; Expert, Average, Novice. The categorization was done based on the experience of each subject. 18 subjects participated in the study where for each expert level there were six subjects. According to the results on average the users have decided that extracted concepts and conceptual relationships are in good quality. Majority (83%) of the users accepted that the suggestions given using web usage patterns were found useful.

Conclusion

The ontology is the backbone of any semantic web application. Developing ontology is a complex, very time consuming and costly task. Therefore, it has become a bottleneck for the rapid growth of semantic web. This fact alone explains the value of this research which semi-automates the process of ontology development.

The main contribution of this research is the concept of using both web author's view point and web user's perspectives in the ontology learning process. Another important contribution of the proposed methodology is , the elimination of the use of web site dependent characteristics in extracting semantics.

This is a promising solution to the current issue that exists in semantic web, which hinders its growth. This methodology could be extended to be used with a large collection of documents which will considerably reduce the cost in terms of time and money in developing semantic web related applications. Another important application of this research is that it can be used for developing cross domain ontologies. Normally ontology is developed for a particular domain. However, through this methodology ontologies could be taken in to a much higher level by developing cross domain ontologies.

References

[1] B. Berendt, A.Hotho, D. Mladenić, M. Spiliopoulou, and G. Stumme,"A Roadmap for Web Mining: From Web to Semantic Web." Springer-Verlag Heidelberg, 2004.

[2] B. Berendt, A. Hotho, and G. Stumme, "Towards semantic web mining," *IN INTERNATIONAL SEMANTIC WEB CONFERENCE(ISWC*, pp. 264–278, 2002.

[3] G. Stumme, A. Hotho, and B. Berendt, "Semantic Web Mining State of the art and future directions," *Journal of web semantic*, pp. pp. 124–143, 2005.

[4] A. Maedche, E. Maedche, and R. Volz, "The Ontology Extraction Maintenance Framework Text-To-Onto," in *In Proceedings of the ICDM'01 Workshop on Integrating Data Mining and Knowledge Management*, 2001.

[5] S. P. Igo and E. Riloff, *Corpus-based Semantic Lexicon Induction with Web-based Corroboration*. .

[6] A. Maedche, E. Maedche, and R. Volz, "The Ontology Extraction Maintenance Framework Text-To-Onto," in *In Proceedings of the ICDM'01 Workshop on Integrating Data Mining and Knowledge Management*, 2001.

[7] E. Maedche and S. Staab, "Ontology Learning for the Semantic Web," *IEEE Intelligent Systems*, vol. 16, pp. 72–79, 2001.

[8] M. Ruiz-casado, E. Alfonseca, and P. Castells, "From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach," in *1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006). Budva*, 2006.

[9] R. W. Cooley, *Web Usage Mining: Discovery and Application of Interestin Patterns from Web Data*. 2000.

[10] E. Mikroyannidis and B. Theodoulidis, *Web Usage Driven Adaptation of the Semantic Web*.

[11] A. D. S. Jayatilaka and G. D. S. . Wimalaratne, “Incorporating web author´s and user´s view in semi automatic ontology creation: A practical approach,” presented at the ICTer, Colombo, Sri Lanka, 2011.

[12] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu, “Integrating Web Usage and Content Mining for More Effective Personalization,” in *IN E-COMMERCE AND WEB TECHNOLOGIES,&QUOT; LECTURE NOTES IN COMPUTER SCIENCE (LNCS) 1875*, 2000.

[13] R. Cooley, B. Mobasher, and J. Srivastava, “Data Preparation for Mining World Wide Web Browsing Patterns,” *KNOWLEDGE AND INFORMATION SYSTEMS*, vol. 1, pp. 5–32, 1999.

[14] M. Grcar, “User profiling: Web usage mining,” presented at the SIKDD multiconference IS,, 2004.

[15] C. Shahabi and F. Banaei-Kashani, “Efficient and anonymous web-usage mining for web personalization,” *INFORMATION PROCESSING AND MANAGEMENT*, vol. 15, pp. 29–35, 2003.

[16] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, J. Riedl, and H. Volume, “Grouplens: Applying collaborative filtering to usenet news,” *Communications of the ACM*, vol. 40, pp. 77–87, 1997.