Available online at www.sciencedirect.com

ScienceDirect

IERI Procedia 7 (2014) 113 – 119

Procedia IERI

www.elsevier.com/locate/procedia

2013 International Conference on Applied Computing, Computer Science, and Computer Engineering

Knowledge-Based Data Mining Using Semantic Web

Sumaiya Kabir[a,b], Shamim Ripon[a]*, Mamunur Rahman[b] and Tanjim Rahman[b]

[a]*Department of Computer Science and Engineering, East West University, Bangladesh*
[b]*Dept. of CSE, Patuakhali Science and Technology University, Bangladesh*

Abstract

Semantic web offers a smarter web service which synchronizes and arranges all the data over web in a disciplined manner. In data mining over web, the accuracy of selecting necessary data according to user demand and pick them for output is considered as a major challenging task over the years. This paper proposes an approach to mapping data over the web 3.0 through ontology and access the required data via an intelligent agent. The agent provides all the searched data related to user query from which user can find desired information. When the user does not have sufficient search parameter, knowledge can be perceived from the information provided by the agent. The derivation of such unknown knowledge from the existing can be achieved by semantic web mining. We present an intelligent agent-based web mining model where users' query is being searched by following existing traditional way, e.g. by Google. The intelligent agent checks the searched data and derives only those are the semantically related to users search parameter. A work-in-progress case study of *University Faculty Information* presented to examine the effectiveness of the proposed model.

Keywords: Semantic Web; Web Mining; Intelligent Agent;

* Corresponding author. Tel.: +88-1928-891-978.
E-mail address: dshr@ewubd.edu.

doi:10.1016/j.ieri.2014.08.018

1. Introduction

Ongoing rapid progress and extensive application of the internet, there is a massive amount of information distributed on the web. The conventional string based search often failed to hit the relevant pages and feedbacks a lot of irrelevant pages from user request. A common problem for a user is that “Everything is on the web, but we just cannot find what we need” [1] is partially true as most of the data over the web is scattered, unstructured, often inconsistent and insufficient. Data sets are not interlinked with each other which makes mining even more difficult to manage.

Discovering unknown knowledge is almost impossible in web2.0, as no relationship is established among data sets making traditional web mining result almost unsatisfactory. For an improved mining, people are now facing toward web3.0. Here, information is presented in a well-defined and structured manner and enable machines and human to work cooperatively. Data in the semantic web is interlinked among each other through ontology which makes effective discovery, mechanization and assimilation possible. These data are machine readable and can be shared and processed by automated tools as well as people.

The semantic web network is a layered architecture [2][3] consists of various levels. In this layered architecture, RDF [4] (Resource Description Framework) and RDF Schema provides a semantic model used to describe the information on the Web and its type. RDF query language SPARQL [5] can be used to query any RDF-based data (i.e., including statements involving RDFS and OWL [6]). The ontology vocabulary layer defines shared knowledge and describes the semantic relationships between various kinds of information. Ontology is considered as the backbone [7][8] for the semantic web architecture as it provides a machine-processable semantics and a sharable domain which can facilitate communication between people and different applications.

The Semantic Web is based on a vision of enriching the Web by machine-processable information. For instance, today’s search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and can thus improve both precision and recall.

Data mining is a process to extract useful and interesting knowledge from large amount of data. Web Mining aims at discovering insights about the meaning of Web resources and their usage. Given the primarily syntactical nature of the data being mined, the discovery of meaning is impossible based on these data only. Therefore, formalizations of the semantics of Web sites and navigation behavior are becoming more and more common. Semantic Web Mining combines Semantic Web and Web Mining. The nature of most data on the Web is so unstructured that they can only be understood by humans, but the amount of data is so huge that they can only be processed efficiently by machines. The Semantic Web addresses the first part of this challenge by trying to make the data (also) machine understandable, while Web Mining addresses the second part by (semi-)automatically extracting the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions. Instead of data mining semantic web enables knowledge mining over web.

Intelligent agent [9] facility enables users to find desired results for all possible related terms with respect to requirements. This paper focuses on how an agent detects all possible entities from ontology during web mining [10] related to a user query request on its own in an automated manner which enables the user to discover unknown knowledge.

In the rest of the paper, in Section 2, we first illustrate our proposed model of semantic web mining and show the steps of how the model can be used. In the following section, we briefly describe our work-in-progress case study of semantic web-based representation to *University Faculty Information*. We briefly describe how an intelligent agent can be used to acquire unknown knowledge with the support of ontology. Finally, we conclude our paper by summarizing our work and outlining our future plan.

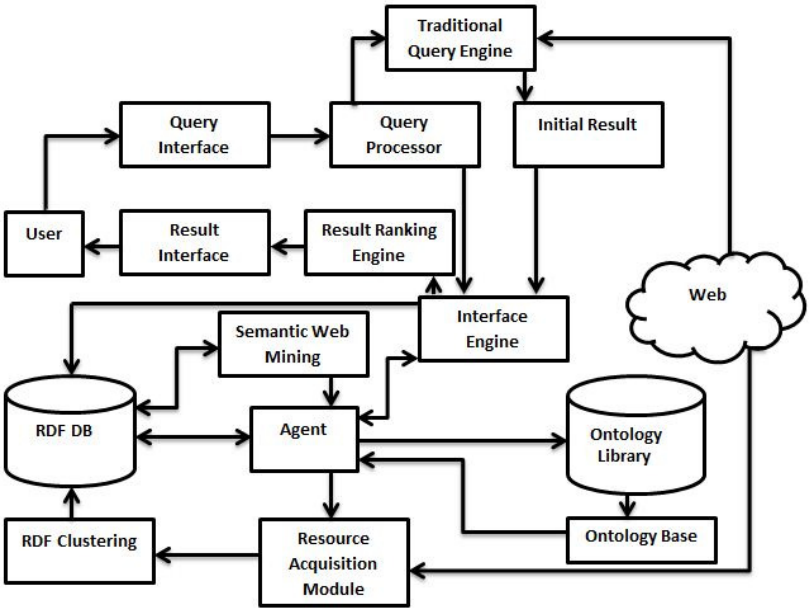

Fig. 1. Proposed web mining model under semantic agent framework

2. Proposed Model

Most of the data over web is unstructured and it is really hard to accumulate them under a common structure. We consider both traditional web mining model and semantic web mining model facilitated by semantic agent for a combination between well-structured semantic network and unstructured real world network situation. Our proposed model is presented in Fig. 1 and it has the following steps.

Step 1: user query request is being sent to query processor through a query interface. The query processor is the subcomponent of the data server that processes user requests.

Step 2: The query processor calls in parallel both traditional query engines and intelligent agent through interface engine with user request as parameters. Interface stop controller enables user to shut down mining immediately, if desired. A query engine is a service that takes a description of a search request, evaluates and executes the request, and returns the results back to the caller. This service acts as an intermediate layer between clients and the underlying data sources by interpreting search requests and shielding the clients from details on how to access the data sources. Traditional query engines return initial results to interface engine and results are sent to RDF database.

Step 3: For agent based searching, an initial ontology should be build and to construct this initial ontology various concepts about the objects of the web need to be gathered together. In most of the cases, specialized clustering algorithm [4] is used to gather data from web. Ontology model merge knowledge of experts [4] in the environment to build initial ontology. The ontology level will be stored in ontology library system [4] for future levels usage.

Step 4: When user request parameters are received by agent from query processor through interface engine, agent checks RDF database. If RDF database contains desired results by caching, agent directly sent results to user through interface engine. On the other hand, agent seeks out all possible relationships between user request and other web entities from ontology library and builds an ontology base with relational entities if desired results are not found in RDF database.

Step 5: Ontology base contains all possible nodes related to user request collected by agent and by acquiring knowledge from ontology base; resource acquisition module collects task related information from the web. But, during acquisition of data from web a crucial problem arises that is arriving of irrelevant information because most of the data over web is unstructured. The total model performance is mostly dependent on these data acquisition performance.

Step 6: Resource nodes of the closest characteristics is detected and collected by resource acquisition module. These nodes are being stored in the RDF database.

Step 7: Semantic web mining module mines the data in RDF database for better output and outputs is being sent to agent.

Step 8: To increase the relevance of result agent performs various filtering process over the outputs of semantic web mining module.

Step 9: In this final step, all the relational results will be sent to interface engine from RDF database by agent. Result ranking engine used for ranking the results and after ranking results will be shown to user by a result interface. The result is given to user exhibits all possible relational aspects from which user could get desired knowledge may be known or unknown. This process is very efficient when users don't have sufficient amount of data parameters to find desired output from web.

3. Ontology-based searching

Ontology level contains all conceptual knowledge about the objects in the field and stores them into ontology library. When a user calls agent with some data parameters, agent starts to search the ontology to find all possible nodes related to user given parameters. This inquiry becomes possible because all the data sets are interlinked with each other and are well defined in semantic web. Agent gives the user a broad range of ability to choose what exactly he/she requires. Thus users feel more comfortable to be facilitated by semantic web agent than web2.0 search engine.

3.1. Case study: University Faculty Information

We design a web of *University Faculty Information* by following semantic web approach. In the case study we consider 40 public and private universities in Bangladesh. For brevity, only Computer Science department related information is considered. We encode generic information about each institution into Protégé [11] using OWL. For each faculty member, we encode information regarding their research projects, funding information, academic and industrial collaborations, publications, etc.

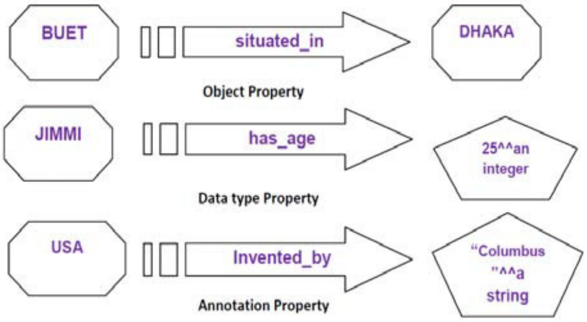

Fig. 2. Various kinds of OWL properties

Individuals, represent objects in the domain in which we are interested. Properties are binary relations on individuals. OWL classes are interpreted as sets that contain individuals. They are described using formal descriptions that state precisely the requirements for membership of the class.

OWL properties represent relationships. There are three main types of properties as in Fig. 2.

- *Object Properties:* Relationships between two individuals.
- *Data type Properties:* Linking an individual to a data literal.
- *Annotation Properties*: Used to add information (metadata | data about data) to classes, individuals and object/data type properties.

There are also various kinds of object properties, namely, *inverse, functional, transitive, symmetric, reflexive*, etc. Along with these properties there are also two restriction properties: *Existential* and *Universal*.

3.2. OWL Ontology Components

Ontology can be accessed by queries and these queries return the relationship between the objects. in RDF Triple format (Subject+ Relationship+ Object). There are two types of user request could be found,

- *Simple Knowledge Acquisition:* User requires total information about a particular faculty member. Relationships with all possible nodes related to Edward will be returned. For example, all the nodes related to Edward are shown in Fig 3.
- *Unknown Knowledge Discovery:* Let us assume that a user wants to know about the relationships between Edward and John. However, in our encoding there is no direct relationship between them. In this situation, agent finds out the closest node related to both Edward and John and the closest node is Jack. Now, agent exhibits all possible relationships between Edward-Jack and Jack-John. From these relationships user would able find out relation between Edward and John (Fig. 4).

4. Conclusion

Main hurdles with the current search engines are related to information overloading. When a search is performed, the most occurring problem is not that too less number of websites are found, but that too many websites are returned. A search resulting in hundreds of thousands websites is more standard than exception.

A lot of these websites are returned because of the simple fact that they contain a word that resembles the search parameter, although the website might not even be relevant to the search query. Another major problem for current search engines is the ambiguous nature of words. For this reason, search engines cannot tell the difference between the different meanings and thus search engines return all websites containing this word, no matter what the meaning is, leaving the user with a big pile of both relevant as well as irrelevant documents.

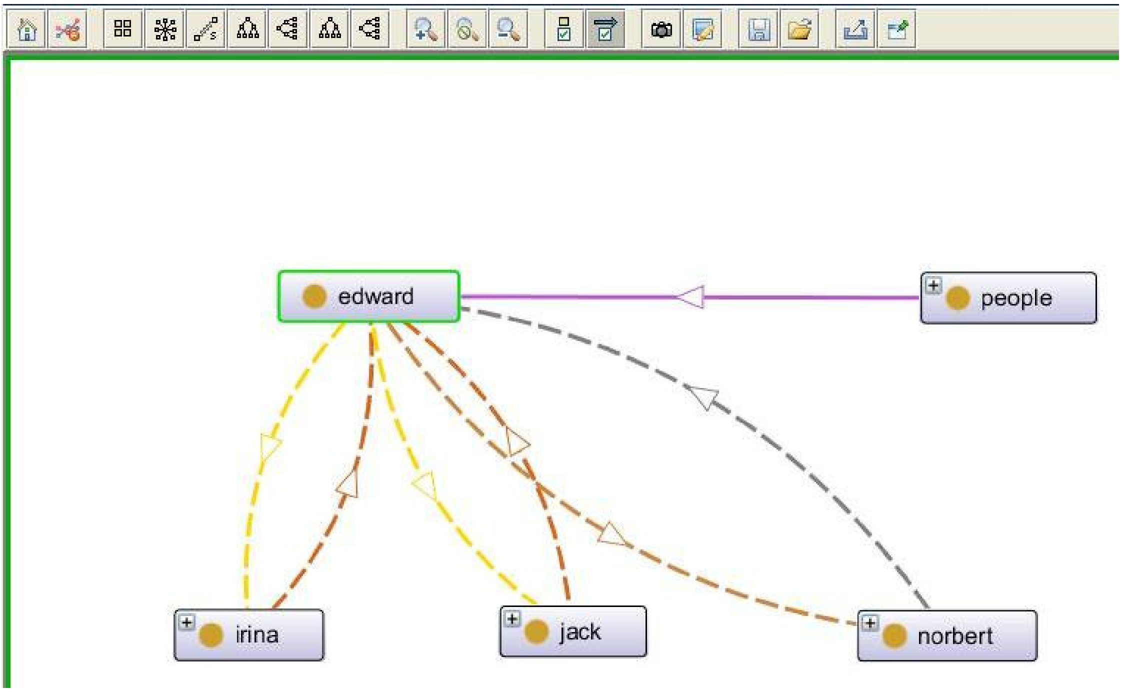

Fig. 3. Simple knowledge acquisition

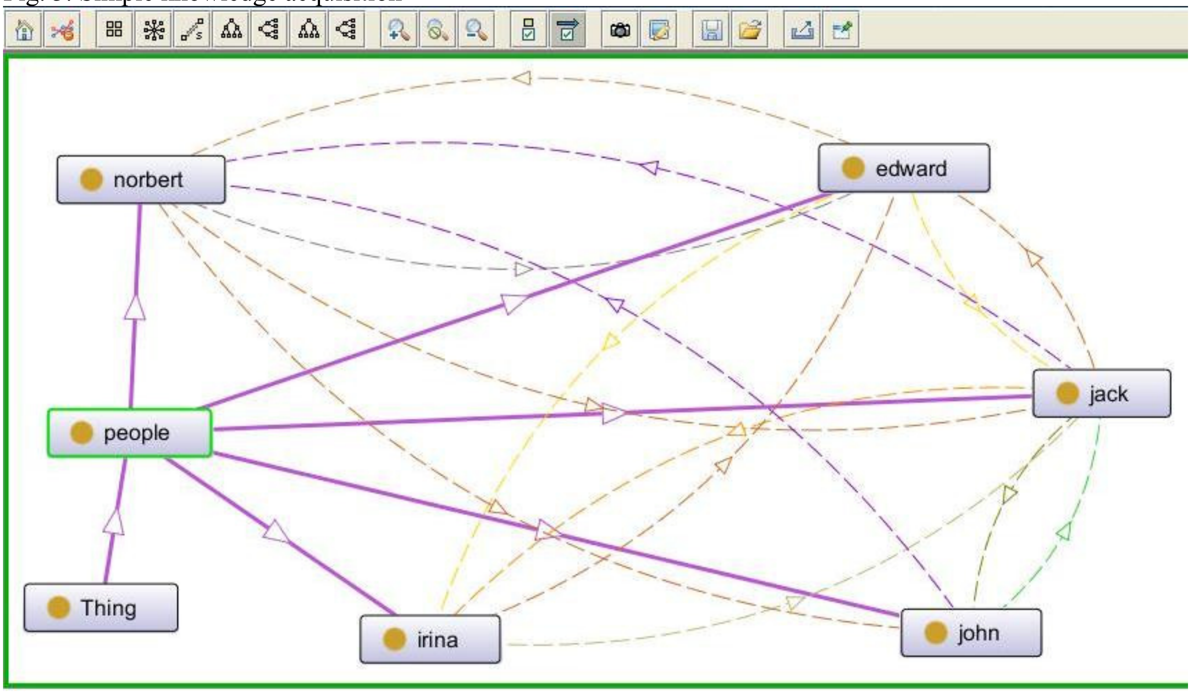

Fig. 4. Unknown knowledge discovery

The Semantic Web envisions a world where agents share and transfer structured knowledge in an open and semi-automatic way. Successfully destination path discovery by agent through ontology as per user request provides various facility such as automation, artificial intelligence, integration, machine to machine communication ability etc. By using these facilities we offer web users knowledge mining instead of data mining. We just developed a model which is facilitated by a single agent and a process to seek out only the desired data from a huge amount of data. In case of multi-agent system, a major concern is the coordination among all kinds of agents.

Our future plan includes dealing with multi-agent environment, their communication, and synchronization. Dealing with the situation where multi-agent learn information from the environment on its own is a critical to knowledge mining.

References

[1] M. Hepp. Semantic Web and Semantic Web Services: Father and Son or Indivisible Twins? IEEE Internet Computing, vol. 10, no. 2, pp. 85-88, March/April, 2006.
[2] W3C Semantic Web, http://www.w3.org/2001/sw/
[3] T. Berners-Lee, Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. 1999,Harper San Francisco.
[4] W. Yong-gui and J. Zhen. Research on semantic Web mining. 2010 International Conference on Computer Design and Applications (ICCDA), vol. 1, pp.67-70, 2010.
[5] A. Seaborne and E. Prud'hommeaux. SPARQL Query Language for RDF, W3C Recommendation, W3C, January, 2008.
[6] I. Horrocks, P. F. Patel-Schneider, F. V. Harmelen. From shiq and rdf to owl: The making of a web ontology language. Journal of Web Semantics 1 (2003) 2003.
[7] A. Grigoris and F. van Harmelen. A Semantic Web Primer. The MIT Press, Cambridge, Massachusetts April 2004.
[8] Semantic Web - XML2000, slide 10". W3C. http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html. [accessed on 13-05-2012]
[9] S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach (3rd ed.). Prentice Hall Press, Upper Saddle River, NJ, USA. 2009
[10] A. Chakravarthy. Mining the semantic web. In: Paper Proceedings of the First AKT Doctoral Colloquium (2005).
[11] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson and M. A. Musen. Creating semantic web contents with protege-2000. In: Protégé-2000. IEEE Intelligent Systems, Vol. 16, No. 2, Mar 2001, pp. 60–71.