Semantic Web Search Using Ontology Based Classification Algorithms: A Survey

R.Umamaheswari 1  Dr.N.Shanthi2

1Assistant Professor, Department of CSE, Gnanamani College of Technology, Nammakal, T.N, India.
2Professor & Dean, Department of CSE, Nandha Engineering College, Erode, TN, India.

ABSTRACT

Now-a-days the web based information is increasing day-by-day. As the number of internet users and the number of web document grows, it is difficult for users to find the documents that are relevant to their particular needs. In Recent years semantic web search is not considering semantic relations between words in traditional Machine Learning algorithms. Previous works on ontology-based semantic web search and web documents classification miss some important issues of dynamic ontology construction and ranking of classified documents. The WorldNet based relations are considered to relate the user queries with the web documents. There is no proper ontology construction is available. Understanding web pages and its contents are very much useful to get better.

This paper analyzes which method is efficient ontology construction for creating semantic relation and classification of web document. This survey Recall, precision rates and similarity improvement methods are studied in detailed manner. This survey paper is used to find out the problems of web information gathering. Ontology plays important role in meaningful and context driven information retrieval. This provides meaningful search and displaying relevant documents. To find out relevant document Similarity measures and classification is used to classify the document according to user query. Finally the page ranking algorithm displays the search results based on page ranking.

1. INTRODUCTION

Semantic web is an extension of the current web describes the evolution of Web as it move towards a more intuitive way of retrieving and presenting search results for searchers. The term "semantic" in this context refers to different meanings, or semantics, of information as this information is presented, and the ongoing effort to make that information more relevant to what searchers are looking for at any given time. The “meanings” based search engine believes that the future of search lies in understanding what the user wants rather than fishing out results with keyword matches. At present, the Semantic Web forming by the RDF [4] documents and OWL documents is one parallel web area, essentially, to the Web forming by HTML documents. Currently, the Semantic Web, (i.e. online documents[8] written in RDF or OWL), is essentially a web universe parallel to the web of HTML documents. Semantic Web documents (SWDs) are characterized by semantic
annotation and meaningful references to other SWDs [5]. Since conventional search engines do not take advantage of these features, a search engine customized for SWDs, especially for ontologies, is needed by human users as well as by software agents and services. At this stage, human users are expected to be semantic web researchers and developers who are interested in accessing, exploring and querying RDF and OWL documents found on the web. This proposed research mainly focus on searching relevant document from web. The available search engines produce search results based on Meta tags. Semantic search engines are very much useful in finding relevant web information.

In this paper, the problems in existing search engines are analysed and also it lists the drawbacks of existing system. Semantic web technologies uses ontology concepts to retrieve meaningful information [3]. Similarity measures are important to find the relevant web document [2]. To find similarity there are many methods available. Here focussing only the ontology based classification algorithm.

2. BACKGROUND

2.1. Web Mining

Web mining is the process of extracting needful information [5] from web pages. Information extraction systems automatically extract precise and exact text fragments from documents. There are three types of web mining. Web content mining, Web usage mining and Web structure mining. Web content mining, is further classified into Hyperlink mining and Document mining. Many of the web pages contain unstructured informations. Finding the relevant document is much difficult in traditional search engines. This paper mainly concentrates on document mining.

2.2. Current Semantic search concepts

Semantic engines are one among the plethora of alternative search options that are now debuting over the web. While many of the new entrants are focusing on improved UIs, clustering or grouping data from engines already there, semantic search is perhaps the most daunting technological approach to the problem. The Semantic Web enriches the text document [4] through the annotations to capture the meaning of the texts. This also helps to improve the description of the online resources. The main concepts of semantic web is Web Search Engine (WSE) and Information Retrieval (IR). Ontology’s plays an important role in semantic web. The data on the Web is not [5] only used to display, but also be understood by the machine so as to enhance the quality of the information services and explore a variety of new, intelligent information services.

2.2.1 Current Semantic Web & Limitations

- Current Web Technologies not suitable for Older search Engines [4]. The documents created by XML are not taken in the traditional search engines.
- Meanings were not included in the search technologies. Semantic tags were not identified by the retrieval process.
- Semantic reasoning was not included.
There are no technologies to combine XML and HTML Tags.

On-line libraries, search engines, and other large document repositories are growing so rapidly that it is difficult and costly to categorize every document manually. In order to deal with these problems, researchers look toward automated methods of working with web documents so that they can be more easily browsed, organized, and cataloged with minimal human intervention. The future works will investigate crawling of dynamic web pages, parallel crawling, and the improved algorithm based on TFIDF algorithm in order to promote the efficiency and precision of traditional web search engine and semantic web search engine.

2.3 Ontology and Classification

2.3.1 Ontology

Many definitions of ontologies have surfaced in the last decade, but the one that in our opinion best characterizes an ontology’s essence is this: “An ontology is a formal, explicit specification of a shared conceptualization.” In this context, conceptualization refers to an abstract model of some phenomenon in the world that identifies that phenomenon’s relevant concepts. Explicit means that the type of concepts used and the constraints on their use are explicitly defined, and formal means that the ontology should be machine understandable.

Ontology is the study of what exists. This is closely related to metaphysics, the study of the nature of reality. In general, ontology deals with the identity of things in the world while metaphysics deals with existential causes such as God or the study of first principles, concepts that underlie all of reality. Autonomy’s technology[7] is built around a Web 2.0 semantic network (based on semantic web concepts) known as an ontology.

Ontology stores search patterns, concepts, and their contexts in an interconnected network. The ultimate goal is to get relevant information from the web. Ontologies provide a flexible framework in which you can create your own words or concepts and link this with other words and with similar or related meaning. Each word can also be linked to its context or domain in which it is used. Ontology[9] is one of the most important concept used in the semantic web infrastructure, and RDF(S) (Resource Description Framework/Schema) and OWL (Web Ontology Languages) are two W3C recommended data representation models which are used to represent ontologies.

In [12] First, Web mining techniques can be applied to help creating the Semantic Web. A backbone of the Semantic Web are ontologies, which at present are often hand-crafted. This is not a scalable solution for a wide-range application of Semantic Web technologies.

The emerging Semantic Web community has proposed ontologies to express knowledge in machine understandable way. The process of building and maintaining ontologies, is known as Ontology Engineering, presents unique challenges.
this paper, they proposed a Semantic Web portal, called OntoKhoj that is designed to simplify the Ontology Engineering process. The methodology in developing OntoKhoj was based on algorithms used for searching, aggregating, ranking and classifying ontologies in Semantic Web. They have implemented the OntoKhoj portal and further validated our system on real ontological data in Semantic Web. We can reuse the existing ontologies to improve the search results. “Two automatic ontology-based document classification methods are described in reference [15] and [16]. Both methods construct ontology by human's help.”

2.4 Semantic Web
As the number of Internet users and accessible Web pages grows, Traditional Web Service search methodology dependant on keywords is time consuming and inefficient. The search results are often inexact. The query often is in unstructured form that is difficult to process as such[2]. One of the main aims of the Semantic Web is lifting the capacity of people and software agents to find documents, information and knowledge in the Web[4]. Ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. It is defined as a "formal, explicit specification of a shared conceptualization"[1]. Basic building blocks of ontology design include:
- Classes or concepts
- Properties of each concept describing various features and attributes of the concept called slots (sometimes called roles or properties).

At present, the Semantic Web forming by the RDF documents and OWL documents is one parallel web area, essentially, to the Web forming by HTML documents[4]. OWL (Web Ontology Language) which is a new formal language is used in the Semantic Web for representing ontologies[6]. Semantic web based on ontology, it processes the unstructured resources into structured information and adds it to the knowledgebase[6]. The ontology mismatch problems include the use of[6]:
- Same terms for different concepts.
- Different terms for the same concepts.
- Semantically similar attributes which have different meanings in their domains.
- Attributes which have different generalization and aggregation level.
- Same attributes, but different data quality requirements, e.g. accuracy.

3. RELATED WORK
3.1. Ontology based Classification
In[10] this paper the construction of ontology applies rules for identification of features to be used for email classification. The training set of emails and used as a part of the feature vectors for an underlying Bayesian classification. In this paper the keeping email community is difficult. They preceded solution for this problem. The ontologies created for emails like characteristics Emails-Agent, Email-Address, Domain, Pole, and message – topics etc.
Protégé tool was used here. More number of email documents should be considered and the result has to be evolved. It deals with only E-mail characteristics. This kind of methodology generally applied for other classification challenges.

In[15] this paper describes the challenges during the implementation with special regard to ontology based classification algorithms, the underlying framework as well as the importance of metadata quality. This is based on the available metadata of files, whereas the classification categories can be found in form of ontology classes. Also a study forecast that the world wide data amount 1800 giga byte in the year 2011. In a decade ago, the problem was inverted, there was not enough digital data to apply information extraction methods. But now-a-days there is plenty of data and we are unable to use this data to extract information and to find important and valuable knowledge in it. The semantic web and ontologies are used to find the gap between several digital data and used knowledge.

Apache TIKA toolkit is used to extract metadata from various files using existing parse libraries. The technique used in this paper is metadata extraction, and after extraction the flexible classes need to be found in order to assign the documents to them. It is possible that one document belongs to multiple classes with the metadata from apache TIKA we can use a PARQL query to find a match for the metadata in the ontology. Using metadata the documents are classified and ontology classes were created. The drawbacks in this approach is that it depends on the quality and quantity of the metadata of the documents. The future work may be extracting the whole information of the document using text running. Using inference engines to find further relationships from the above paper.

In[14] the above paper the process of web document classification involves calculating similarities between documents and categories by using the information extracted from them. In recent years Ontology based web document classification method is introduced to solve the problem of classifier training and not considering semantics relations between words in traditional machine learning algorithms. After classification the simple ranking algorithm is used to display the web pages. The Precision and Recall performance were increased compared to the KNN (K-Nearest Neighbour) and SVM (Support Vector Machine) algorithms. Weighted Word set were taken for calculating and display the results. In future weighted word sense will be considered.

Here [6] ontology web language plays important to derive hidden information of ontology a number of owl reasoners have been introduced. In this paper an enhanced method of optimizing the ontology reasoning. The result of the experiment was that the enhanced search method increases performance improvement 30%. Ontology is a formal, explicit specification of a shared conceptualization of a domain of interest [1]. Therefore, many ontology building and reasoning systems can be used to represent taxonomic and conceptual knowledge of a problem domain in a structured and well formed way. Future work will be increasing.
search performance. They checked the result using kolala.owl, pizza.owl, miniEconomy .owl and miniTransport.owl ontologies.

Research project selection [9] is an important task for government and private research funding agencies. When a large number of research proposals are received, it is common to group them according to their similarities in research disciplines. Text mining methods are used to classify the proposals based on the category.

Ontologies are created according the different areas. For classification last 5 years proposals were taken as a training set and documents are classified. Research update is annual update. Reference [17] represents an ontology based classification method which can process Chinese documents, the approach establishes an ontology model based on Hownet knowledge based and its method, and classifies documents using a simple ontology relevance calculating algorithm. Reference [14] Earth Movers Distance algorithm is used to calculate similarity distance between web documents using weighted graphs. To find the semantic meaning WordNet lexical database is used as in [18].

Future work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically. Also, there is a need to empirically compare the results of manual classification to text-mining classification. Finally, the method can be expanded to help in finding a better match between proposals and reviewers.

4. PROPOSED RESEARCH

Referring the papers semantic web is the future of next Internet technologies. Instead of retrieving information based on Metadata semantic relations provides efficient Information Retrieval (IR) with ontologies. Problems in traditional search engine as follow:

- No proper Semantic relations.
- Dynamic ontology construction not available.
- Traditional classification method requires training module to classify the data. Also classifiers not consider semantic relations.
- Ontology classification available only for unstructured information.
- Word sense sets are not analysed.

4.1 The proposed architecture of semantic web search engine

There are 6 main Modules used to get relevant information.

1. URL Sets
2. Query processing
3. Collecting Web documents
4. Dynamic ontology construction and classification
5. Calculating similarity score
6. Ranking the documents
Fig 1 : Proposed Semantic Web search Architecture

In this architecture users are giving the search information. Based on the search information the URL sets module collects URL sets from the web dynamically and the database is stored. Also first proposed plan have experiment with URL sets collected from repositories. From the URLs sets the documents are extracted. The web documents are classified using ontology’s using similarity score. The similarity score is calculated by comparing various similarity score methods available in [2]. The Earth Movers Distance(EMD) method will be researched and Finally best method is selected to calculate the Similarity score.

In[14] simple ranking algorithm based on similarity score is used. In proposed system the various ranking algorithm is analysed to sort the web documents to achieve better performance.

5. CONCLUSION

In this paper the brief survey in existing semantic web search based on ontology classification algorithm is made. The various issues in the above method is analysed based on users perceptions and to provide low precision and high recall through experimental results. In the future, the research work will focus on the deeper research in the field of Semantic web search, with the purpose of to overcome current situation of the field and promote the development of semantic web search engine using ontology based classification.

6. REFERENCES

[1] International journal of web & semantic technology (ijwest) vol.2, no.1, January 2011 doi: 10.5121/ijwest.2011.2103 34 intelligent semantic web search engines: a brief survey g.madhu(madhu_g@vnrvjiet.in), Dr.a.govardhan and dr.t.v.rajinikanth
[2] Unsupervised semantic similarity computation between terms using web documents elias iosif, student member, ieee, and alexandros potamianos, senior member, ieee transactions on knowledge and data engineering, vol. 22, no. 11, november 2010
[3]The research of search engine based on semantic web yi jin, zhuying lin, hongwei lin school of mathematics and computer science,guizhou normal university, guiyang, guizhou, 550001, china, international symposium on intelligent information technology application workshops
[4] Research on semantic web mining, wang yong-gui1dept of software liaoning technical university, jia zhen2 dept of electronics and information engineering,liaoning technical university huludao, liaoning, china, 2010
international conference on computer design and applications (iccda 2010)
kimjemins@hotmail.com, kwonshzzzang@naver.com, park@ssu.ac.kr
[8]Ontology-based classification of email kazem taghva, julie borsack, jeffrey coombs, allen condit, steve lumos, and tom nartker information science research institute university of nevada, las vegas
taghva@isri.unlv.edu
[9] Stumme, hotho, and berendt: semantic web mining state of the art and future directions gerd stumme, andreas hotho, bettina berendt
[10] Ontokhoj: a semantic web portal for ontology searching, ranking and classification chintan patel, kaustubh supekar, yugyung lee, e.k. park school of computing and engineering university of missouri-kansas city {copdk4, kss2r6, leeyu, ekpark@umkc.edu}
[11] Ontology-based classification of unstructured information, stefan burger eastern michigan university department of computer science ypsilanti, michigan, usa sburger3@emich.edubernd
stieger abbcorporate research center industrial software and applications ladenburg, germany
bernd.stieger@de.abb.com, 978-1-4244-7571-1/10/$26.00 ©2010 ieee 254
[12] Ontology-based automatic classification and ranking for web documents jun fang, lei guo, xiaodong wang and ning yang control and networks laboratory, school of automation, northwestern polytechnical university, china {fangjun_nwpu, lguo, hellowxd, yangning@mail.nwpu.edu.cn