Ontological Paradigm for Focused Crawling based on Lexical Analysis

Nidhi Sharma
Department of Computer Science
Echelon Institute of Technology
Faridabad, India

Atul Srivastava
Department of Computer Science
Echelon Institute of Technology
Faridabad, India

ABSTRACT

The semantic web is a synergetic movement led by International standards body, the WWW Consortium (W3C).It aims at converting the current web dominated by unstructured and semi structured documents into a “web of data”. Here two techniques of semantic web crawling are reviewed, one is ontology based and other is based on Lexical database .For this, architecture has been proposed which is a combination of above two techniques. The future of WWW is semantic web where Ontology and Lexical database are used for effective and fast searching by the web crawler. It is used for Information retrieval and question answering system. Ontology is a formal designation of shared approach it is basically approach of entities and their attributes.

General Terms:

Information Extraction, Semantic Matching.

Keywords

Crawler, Semantic web, Ontology, Lexical database

1. INTRODUCTION

Defining the Semantic Web is a complicated task. It is the next generation of the web. It is a set of languages and standards. It has a strong logic and reasoning component. Semantic search systems consider various points including context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results [1]. The new system differs from other representation systems in that it is based on a more sophisticated semantic representation of information, aims to go well beyond the document level, and designed to be understood and processed by machine. A common theme underlying these three features, i.e., turning documents into meaningful interchangeable data, reflects a rising use expectation nurtured by modern technology and, at the same time, presents a unique challenge for its enabling technologies.

In this paper, two emerging trends of research in the Semantic Web space are addressed and begin by presenting two techniques for searching content on the Semantic Basis: an ontology domain and Lexical database. The proposed architectural framework which is combination of the above two techniques provides the effective and significant results.

1.1 CRAWLER

A program that searches the World Wide Web, typically in order to create an index of data .Crawlers apparently gained the name because they crawl through a site a page at a time, following the links to other pages on the site until all pages have been read.

1.2 THE SEMANTIC WEB

"The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries” It is a source to retrieve information from the web (using the web spiders from RDF files) and access the data through Semantic Web Agents or Semantic Web Services[2]

The basic concept of having certain semantic information on the Web pages that can be used in order to solve typical obstacle of Information extraction and Question Answering .It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users.. In this manner the Semantic Web relies on two basic components, ontology and lexical database. It relies on ontology in order to interpret the textual content of a resource and uses the Lexical database to extract the metadata from data extractor to provide sense and logic to the data Download priorities are assigned to pages by applying semantic similarity criteria for computing page-to-topic relevance: a page and the topic can be relevant if they share conceptually (but not necessarily lexically) similar terms. Conceptual similarity between terms is defined using ontologies [3] [4] [5].

1.3 ONTOLOGY

The goal of Ontology is to represent the collective knowledge intended for the use of a group. Ideally the Ontology captures knowledge independently of its use and in a way that can be shared, but practically different tasks and uses call for different representations of the knowledge in Ontology.

Ontology has a richer internal structure as it includes relations and constraints between the concepts. It claims to represent a certain consensus about the knowledge in the Domain [6]. This collection is among the intended users of the knowledge, e.g. players using a game Ontology regarding a certain game. “Ontology is domain specific”

Learning of Ontology as shown in fig. 1 which starts with the distillation of specific keywords related to query from domain Ontology, then prune and permeate the data, obligate this procedure on web documents and then store and reprocess the data for lateral purposes. The domain Ontology used by the system may be generated internally by an Ontology Generator component.

1.4 LEXICAL DATABASE

Lexical database is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas. Its design is inspired by current psycholinguistic and computational theories of human lexical memory [7]. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different relations link the sets. Most commonly used Lexical database are Word Net, concept net and YAGO [8, 9, 10].

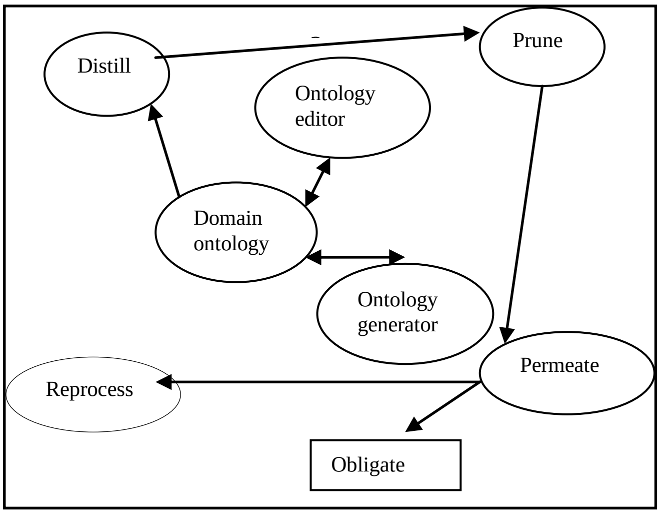

Fig. 1 Ontology learning process

Word Net model is Lexical database which is composed by three main classes: Synset, Word sense and Word. Word Net [11] toolkit is widely used for the English language .it group English words into sets of synonyms called Synset and provides semantic relationships between them includes a taxonomy. The first two are comprised into four:

(Class Hierrarchy in Word Net)

Synset

- Adjective Synset
- Adjective Satellite Synset
- Adverb Synset
- Noun Synset
- Verb Synset

Word Sense

- Adjective Word sense
- Adjective Satellite Word sense
- Adverb Word Sense
- Noun Word Sense
- Verb Word sense
- Word
- Collocation

Lexical types subsets:- Noun, verb, adverb and adjective. The only subset of word is collocation used to represent words that have hyphens or underscores in them.

1.5 Comparative Study Of Ontology And Lexical Database

Here, Comparisons of these two techniques are discussed on the basis of their functionality, importance, their types and goal.

Table 1. Comparisons of Ontology and Lexical Database

	ONTOLOGY	**LEXICAL DATABASE**
Functionality	Provides declarative representation of knowledge relevant to a particular domain.	Provides sense of a given word which is called as synset.
Importance	It provides Well-defined meaning of the information.	It provides the semantic relationship.
Semantic Web usage	Semantic matching based on Ontology gives the better result.	It creates semantic Knowledge base of words that are retrieved from metadata.
Types	Widely used Lexical database are Conceptnet, Wordnet and YAGO	Widely used Ontology based information extraction are Text-to-Onto and Ontox.
Goal	It provides foundations to build other Ontology and remove the problem of ambiguity.	It provides significant and effective results.

In semantic web, combinations of these two techniques enhance the performance of the web and provides unambiguous and domain specific results.

2. RELATED WORK

The variety of search engines such as Google, Bing and similar other search engines are used for crawling the web page or documents from the WWW. Earlier in Semantic web tokenization algorithm, indexing the documents, page Ranking, Focused crawling and Ontology based Page relevance algorithm are used.

2.1 FOCUSED CRAWLER

The goal of a focused crawler is to selectively seek out pages that are relevant to ad-hoc queries. This crawler advertise semantic web data, by use of some fact finding to rate pages which is appropriate for the user and related to the topic, so that crawler should not pursued the inappropriate web pages.

2.2 PAGE RELEVANCE

Location of a word or phrase is a factor that most search engines will consider relevant. Pages with keywords appearing in the title are assumed to be more relevant than others to the topic, as well as pages with keywords that appear near the top the page.

Frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in a web page. Those with a higher frequency are often deemed more relevant than other web pages.

Proximity is another indicator of relevance. If those words are found close together in a document, that document is assigned a higher weight than one in which the words appear scattered farther apart. When a user searches for a number of words, and in the document in which the words are found close together, it's more likely that they are being used in the same context as the user meant.

2.3 ONTOLOGY BASED PAGE RELEVANCE ALGORITHM

An Ontology use to extract terms and concepts from plain text using different terminology extraction. If we used this algorithm for page relevance, user gets the relevant pages as per his desires or requirement by the use of various measures on Ontology graph.

In these Abstract shows the critical issue which reveals that for example considers a situation or result shown in form of Record1 and Record2 both had result about APPLE, but these results were not sufficient to show which link the user must pursue to get exactly the same result which is desired by the user. There may be different meaning of Keyword 'apple' like Company and fruit but Record1 and Record2 do not specify this. Here, User faces the problem of ambiguity.

Our approach is the combination of two techniques for the effective and precise results. This can be achieved by the use of Ontology based and Lexical database on semantic web for extracting the relevant documents desired by the user. Semantic Knowledge Base is used for storing and reuses the documents for further extraction.

3. PROPOSED WORK

The Proposed methodology has two main modules such as Ontology domain and Lexical database which stores the data in Semantic Knowledge Base after crawl the web pages from crawler. In this paper we optimize the database size. Due to optimized size of database, user got the higher precision rate.

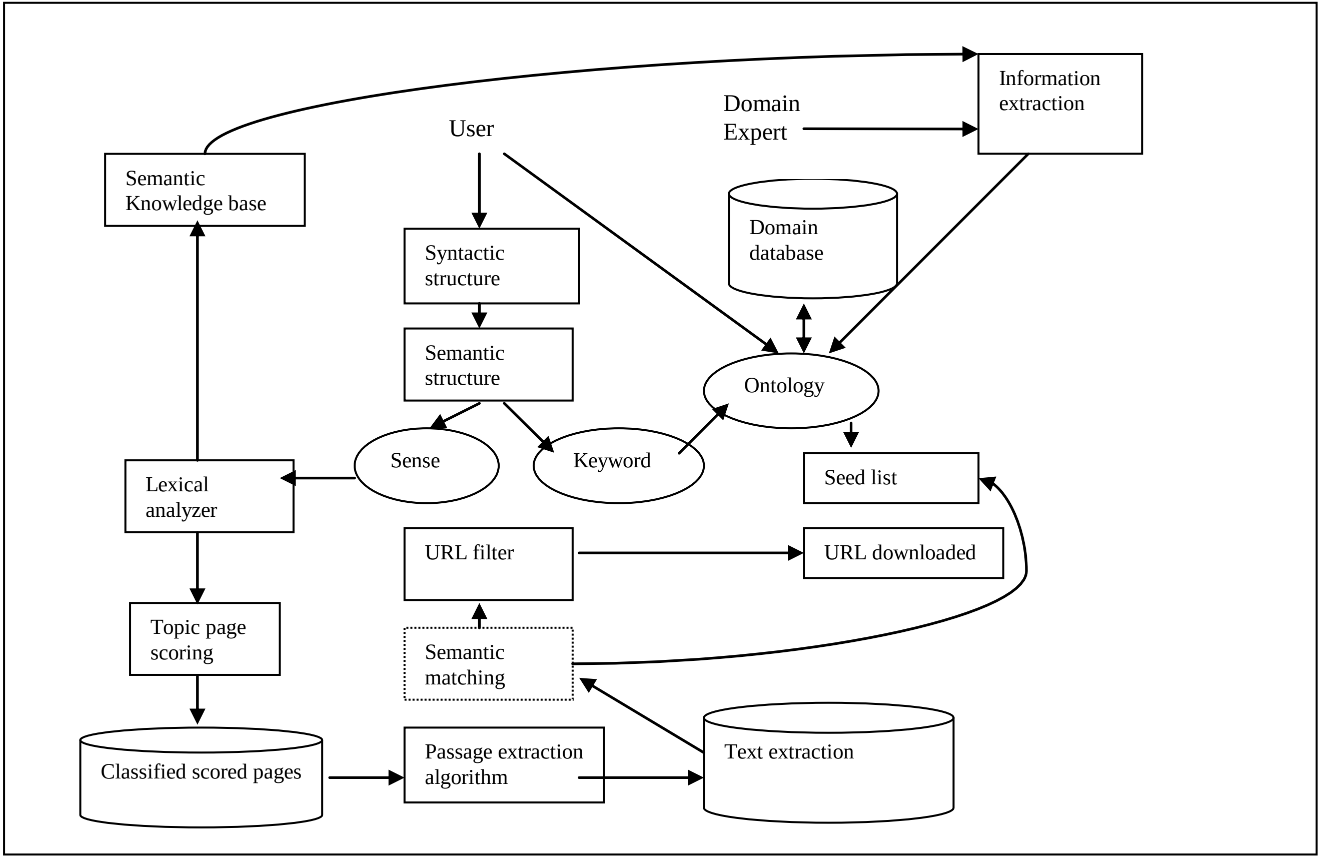

Fig 2.Architectural Framework

3.1 ARCHITECTURAL DIAGRAM

Crawler is used to crawls the web pages to extract the metadata. According to the fig 2, User enters the query which is represented in syntactic structure form and it also gives the contextual information to Ontology simultaneously, which check for the relevant context in the Domain database. Then query is processed by semantic web which is extension of traditional web and further divided into two segments one is Sense and other is Keyword. Keywords are processed by Ontology which is interconnected with Domain database for the relevant documents. Domain Expert may also be involved in the information extraction process in some systems that operate in a semi automatic manner. Ontology deals with the information extraction module to extract the text related information. The output of Ontology base information extraction process is stored in domain database. An approach such as SOR [12] can be used to store Ontology in the database.

This database gives feedback to Ontology for further updating of domain database based on Ontology for the related or relevant documents as well as it sends the specific documents to Lexical Analyzer and Sense is also analyzed by the Lexical database such as Word Net to find out the meaningful and unambiguous final result and display it to the user. Metadata information is stored in semantic knowledge base for further extraction in future. Word Net lexical database is used to generate the semantically related terms, the so called thematic terms. Topic relevance scoring maps the 'pages' keywords to their corresponding Ontology nodes in order to compute an appropriate Ontology topic for representing the pages thematic content .Then pages are classified .Passage extraction algorithm is employed to extract from the pages contents that is semantically closest to the identified topics. Apply the Wu and Palmer [13] similarity measure, which computes the degree to which passage terms semantically relate to the Ontology concepts that represent focused categories. Through semantic matching [14, 15] user get the highest relevant page according to topic, then user get the desired result such as relevant Url is downloaded. Information Extraction module used to extract information related to Ontology such as instances and property values.

Ontology and Lexical Analysis techniques provide the simple and good results. It provides the index freshness, accurate, precise and unambiguous results.

This overall approach gives the best, efficient, parsed and precise and significant results.

Table 2. Fundamentals of proposed architecture with their pros and cons

FUNDAMENTAL	FUNCTIONALITY	PROS	CONS
Ontology database	Ontology is used for Matching.	Provides background knowledge that allows non-experts to query from their point of view.	Great level of abstraction, difficulty to maintain a consistent logic.
URL Crawler	Extracts the relevant URLs from the list.	Efficient	Download irrelevant URLs also
Lexical database	Provides the semantic relationship and retrieve metadata.	Gives sense to data	Complex
Semantic matching	Matches process based on certain parameters like common definition matching and uses ontology for page relevance computations.	Filter the modules.	Ambiguous
Semantic knowledge base	Gives most relevant results.	Store and reuse	Expensive

4. EXPERIMENTAL ANALYSIS

In focused crawler the relevance score reflects the importance of page is, given the topic of the crawl. First of all, start an unfocused crawler from a set of seed URLs, and within the first few hundred page fetches, it was completely lost in web terrain having nothing to do with bicycling. The precision rate is low as compared to focused crawler.
It is crucial that the precision rate of the focused crawler be high, otherwise it would be easier to crawl the whole web and bucket the results into topics. Initially, starting from the list of seed URLs, kept up a healthy precision rate, collecting relevant pages almost half the time by setting the application limit to 5 and Endurance limit is 3. Within two hours on a small desktop, User crawled over 5000 pages relevant to bicycling. User did not require any dependence on a general crawl. Finally, he/she got the desired result. Within analysis and design, two main areas of application are identified: Semantic Crawlers: First, precision Rate and second is Recall graph. Semantic Crawler based on Synonym, Hypernym & Hyponym similarity.

In Information Retrieval, Precision and Recall are the two most widely used metrics for performance measurement. Precision shows the number of correctly identified items as a proportion of the total number of items identified while recall shows the number of correctly identified items as a proportion of the total number of correct items available. Using {Relevant} and {Retrieved} to denote the sets of relevant and retrieved documents respectively, precision and recall are often represented using the following formulae [16, 17]. They are related to the usage of these measures in information retrieval.

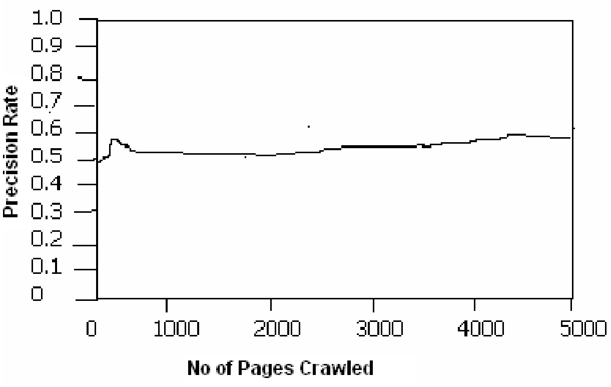

Fig 3.Experimental Outcomes

$$\text{Precision} = \frac{|\,\{\text{Relevant}\} \cap \{\text{Retrieved}\}\,|}{|\,\{\text{Retrieved}\}\,|}$$

$$\text{Recall} = \frac{|\,\{\text{Relevant}\} \cap \{\text{Retrieved}\}\,|}{|\,\{\text{Relevant}\}\,|}$$

5. CONCLUSION AND FUTURE WORK

Architecture for Ontology and Lexical database based semantic web crawler populates the semantic Knowledge base with most relevant and meaningful resources as desired by the user. It gives the efficient, unambiguous, effective and precise results. This architecture guides the URL crawler for extracting relevant information which provides the wider scope for better search engine. Query optimization is done in this paper. Future directions are to improve the flaw of complexity and it can be improve, if there is a combination of other semantic web techniques are used like Description logic and information retrieval algorithms. Further, Ontology based Information extraction can be used to evaluate the quality of Ontology.

6. REFERENCES

[1] John, Tony (March 15, 2012). "What is Semantic Search?". *Techulator*. Retrieved July 13, 2012

[2] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu, "Intelligent crawling on the world wide web with arbitrary predicates," in World Wide Web, 2001, pp. 96–105. .Available:iteseer.ist.psu.edu/aggarwal01intelligent.html.

[3] M. Ehrig and A. Maedche, "Ontology-focused crawling of web documents," in Proc. of the Symposium on Applied Computing, March,Florida, USA, 2003.

[4] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34{43, 2001.

[5] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G.M., Milios, E. "Semantic Similarity" International Journal on Semantic Web and Information Systems (IJSWIS) .

[6] Debajyoti, Arup Biswas, Sukanta "A New Approach to Design Domain Specific Ontology Based Web Crawler", 10th International Conference on Information Technology – 2007 IEEE.

[7] Fabian M. Suchanek , Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (WWW '07).

[8] Miller, G. A. "WordNet : A Lexical Database for English," Communications of the ACM (Vol. 38, No.11), 1995, pp. 39-41.

[9] Fellbaum, C. WordNet: An Electronic Lexical Database, Cambridge, MA: MIT, 1998.

[10] Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, To Appear. Volume 22, forthcoming issue. Kluwer Academic Publishers.

[11] Methods in "Word Net and their Application to Information Retrieval on the Web" 7th ACM

[12] J. Lu, L. Ma, L. Zhang, C. Wang, Y. Pan, and Y. Yu, SOR: a practical system for Ontology storage, reasoning and search

[13] Wu X. and Palmer M.: Web semantics and lexical selection. In the 32nd ACL Meeting, (1994)

[14] Semantic Similarity" International Journal on Semantic Web and Information Systems (IJSWIS) .

[15] Special Issue of Multimedia Semantics, 2006, Vol. 3 July/September, No.3, pp. 55-73.

[16] J. Han and M. Kamber, Data Mining: Concepts and Techniques 2nd Edition (Morgan Kaufmann, San Francisco, CA, 2006) 616-617.

[17] Precision and Recall (2009). Available at: http://en.wikipedia.org/wiki/Precision_and_recall (accessed 25 June 2009).