Enhancing the performance of web Focused CRAWLer using ontology

Prasant Singh Yadav, Mrs Mala Kalra, Dr. K.P Yadav
M.E Scholar Department of Computer Science & Engineering, NITTTR, Chandigarh, India
Assistant Professor Department of Computer Science & Engineering, NITTTR,Chandigarh, India
Professor Department of Computer Science & Engineering,SIET,Ghaziabad,U.P,India
Pdesire82@gmail.com,malakalra2004@yahoo.co.in,karunesh_732@gmail.com

Abstract: The enormous growth of the World Wide Web in the recent years has made it important to perform resources discovery efficiently. The rapid growth of World Wide Web poses (Doubles in size approximately every eight months) unprecedented scaling challenges for general purpose crawler and search engine. Finding useful information from the web which has a large and distributed structure required efficient search strategies. As ontology plays an important role in providing controlled vocabulary of concepts, each with an explicitly defined and machine process able semantics. In this paper ,we propose the novel concept of intelligent crawling of Ontology based content focused crawling , the new approach that analyses it crawl boundary to find the links that are likely to be the most relevant for the crawl while a boundary irrelevant region of the web. Through our new focused crawling technique we solve the polysemy (refer to word with multiple meaning) and synonymy (refers to multiple word having the same meaning) semantic net problem. Also instead of searching in the whole web, our proposed technique will search in the ontology build by us that is updated periodically after a very short interval than instead of displaying all the information that is not related to the user need, we will display only relevant and related information. Our purposed work give us two fold benefit , firstly only focused result are retrieved which reduce the number of results entreated and secondly, due to focused searching irrelevant result are pruned which reduce the time.

Keywords: Ontology, Focused Crawler, Search Engine,

I. Introduction

A search engine is an information retrieval system designed to help to minimize the time required to find information over the vast Web of hyperlinked documents. It provides a user interface that enables the users to specify criteria about an item of interest and searches the same from locally maintained databases. The criteria are referred to as a search query. In the case of text search engines, the search query is typically expressed as a set of words that identify the desired concept that one or more documents may As compared to traditional document collections which reside in physical warehouses such as the college's library; the information available on WWW is distributed over the Internet. In fact, this huge repository is growing rapidly without any geographical constraints.

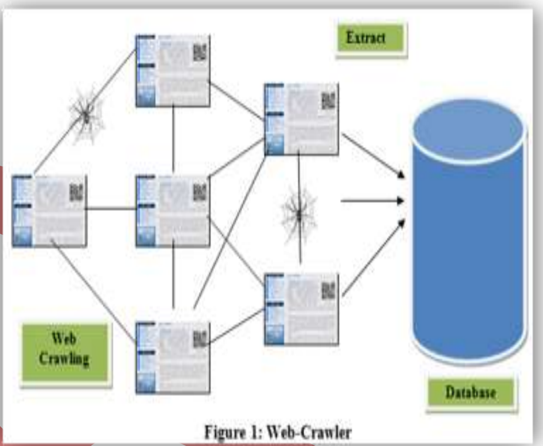

Figure 1: Web-Crawler

Therefore, a component used crawler is employed by the search engine which visits the Web pages, collects them and categorizes them. The crawler retrieves web pages commonly for use by a search engine. It traverses the web by downloading the documents and following embedded links from page to page.Formally, crawlers may be defined as "Software programs that traverse the World Wide Web information space by following the hypertext links extracted from hypertext documents".

II. Focused Crawler

A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics. They attempt to download pages that are similar to each other. The concepts of topical and focused crawling were first introduced by Chakrabarti et.al [3, 9]. The main problem in focused crawling is that in the context of a Web crawler, we would like to be able to predict the similarity of the text of a given page to the query before actually downloading the page. A possible predictor is the anchor text of links; this was the approach taken by Pinkerton in the first web crawler of the early days of the Web. The performance of a focused crawling depends mostly on the richness of links in the specific topic being searched, and a focused crawling usually relies on a general Web search engine for providing starting points.

Ontology always includes a vocabulary of representational concept labels to describe a shared domain. These concept labels are usually called terms (lexical references) and are associated with entities (non lexical referents – the concepts) in the universe of discourse. Formal axioms are also introduced to constrain their interpretation and well-formed use. Ontology is in principle a formalization of a shared understanding of a domain that is agreed upon by a number of agents described by Spyns P. et al [25]. In order for this domain knowledge to be shared amongst agents, they must have a shared understanding of the domain and therefore, agreement must exist on the topics about which to communicate. In other words, in order to facilitate meaningful communication, an agent must commit to the semantics of the terms and relationships in the common ontology declared by J.M Park et al [22]. This includes axioms about properties of objects and how they are related, also called the semantic relationships of the ontology.

III. Related work

Ontology has been used to improve the effectiveness of focused crawling. Hiep Phuc loung et al [29] Ontology based Crawling and A. Ardo [8] web crawler estimates the semantic content of the link of the URL in a given set of documents based on the domain dependent ontology, which in turn strengthens the metric that is used for prioritizing the URL queue. The link representing concepts in the ontology knowledge path is given higher priority. Hong –Wei Hao et al [30] considers an ontology-based algorithm for page relevance computation. After preprocessing, entities (words occurring in the ontology) are extracted from the page and counted. Relevance of the page with regard to user selected entities of interest is then computed by using several measures on ontology graph (e.g. direct match, taxonomic and more complex relationships). The harvest rate is improved compared to the baseline focused crawler (that decides on page relevance by a simple binary keyword match).

Chakrabarti et al [9] proposed that the next generation of the Semantic Web focuses on supporting a better cooperation between humans and machines. In this approach, ontology plays an important role as a backbone for providing and accessing knowledge sources. Since manual building of ontology is costly, time-consuming, error-prone and inflexible to change, it is hoped that an automated process will result in a better ontology construction and create ontology that better match a specific application represented by A.Maeche et al [13]. Ontology learning approaches can be distinguished by the type of input used for learning, e.g., they can learn from text, from a dictionary, from a knowledge base, from a semi structured schemata, or from relational schemata described in A.Gomez, M. Samsfard [10, 16]. Currently, few projects attempt to support the entire ontology learning process including automated support for tasks such as retrieving documents, classifying, filtering and extracting relevant information for the ontology enrichment. Most existing approaches for ontology learning require a large number of input documents for accurate results as in B. Omelayenko [15]. With the enormous growth of the Web, it is important to develop document discovery mechanisms based on intelligent techniques such as focused crawling in T.joachims [11] to make this process easier for a new domain. Focused crawlers go a step further than classic crawlers in order to be able to quickly collect Web pages about a particular topic or domain of the Web.

Gómez-Pérez et al. [10] presents a good summary of several ontology learning projects that are concerned with knowledge acquisition from a variety of sources such as text documents, dictionaries, knowledge bases, relation schemas, semi-structured data, etc. Many of these existing approaches employ ontology learning from text documents, although only a few deal with ontology enrichment from documents collected from the Web. Omelayenko [15] has discussed the applicability of machine learning algorithms to learning of ontology from Web documents and also surveys the current ontology learning and other closely related approaches.

Similar to our approach, J.M park et al in [22] introduces an ontology learning framework for the Semantic Web which proceeds through ontology import, extraction, pruning, refinement, and evaluation giving the ontology engineers a wealth of coordinated tools for ontology modeling. However, they do not mention any automated support to collect the domain documents from the Web or how to automatically identify domain relevant documents needed by the ontology learning process. In another approach similar to ours, S.T. Dumais et al [17] has presents an automatic method to enrich very large ontology, e.g., World Net that uses documents retrieved from the Web. But, they do not apply any filtering techniques to verify that the retrieved documents are truly on-topic.

Many ontology learning approaches require a large collection of input documents in order to enrich the existing ontology as in B. Omelayenko [15]. A common way to get these documents from the Web is to use general purpose crawlers and search engines, but this approach faces problems with scalability due to the rapid growth of the Web. In contrast, focused crawlers overcome this drawback, i.e., they yield good recall as well as good precision, by restricting themselves to a limited domain [18].Devashis hati et al [34] describe a new hypertext resource discovery system with the purpose of selectively seeking out pages that are relevant to a pre-defined set of topics. Ester et al [18] also introduce a generic framework for focused crawling consisting of two major components: (i) specification of the user interest and measuring the resulting relevance of a given web page; and (ii) a crawling strategy. In order to improve accuracy of the learned ontology, the documents retrieved by focused crawlers may need to be automatically filtered by using some text classification technique such as Support Vector Machines (SVM), k-Nearest Neighbors, Linear Least-Squares Fit, TF-IDF, etc. A thorough survey and comparison of such methods and their complexity is presented in J.Qin [20] and C.C aggrwal et al [1] conclude that SVM to be most accurate for text classification and fast training. M.Ehrig et al [18] and T. Joachims [11] described SVM as a machine learning model that finds an optimal hyper plane to separate text classification and fast training and then classifies data into one of two classes based on the side on which they are located.

We also adopt the meta-search method proposed by J.Qin [20] in our framework. Other works more related to ours mostly adopt a certain semantic model in crawling. S.Charkrabarti et al [9] uses thesaurus to process predefined documents associated with the specified topic. S.M Pahlevi et al [19] combine the taxonomy-based search engines and a machine learning technique for adaptive Web search. Ester et al [18] uses a complex ontology and associated instance elements to build the focused crawler. Hong-wei Hao[33] also defines the topic focus as an ontology, which is used for automated subject classification. J.Graupmann et al [21] builds a search engine which crawl semantic markups in HTML, XML, etc.

IV. Present Problem

Ontology provides a base framework for knowledge representation, and the methodology of ontology construction is one of the most important research topics in the ontology community. Many methodologies have been proposed, and some of them have been along with constructing engineering ontology. However, the previous methodologies are mostly top-down approaches which do not maximize the benefits of bottom-up approaches. There are few bottom-up approaches, but they do not utilize the full resources of knowledge such as engineering documents.

A critical look on the available literature reveals that the existing work needs to include the following issues:

There is a need of search engine which cover the two major issues of information retrieval i.e. Polysemy and Synonymy that to simultaneously. Polysemy refers to words with multiple meanings, i.e. how the same phonological form (word) has different semantic mappings (meanings). If the two meanings are unrelated, as in the word pen meaning both writing instrument and enclosure, they are considered homonyms. Synonymy refers to multiple words having the same meaning. As the name implies, synonyms are words that mean the same or have similar meanings in context. Synonyms are used in a variety of situations not only for variety, but to express thoughts or ideas in another, often more emphatic manner.

To make web searching specific and fast, an appropriate ontology construction plays the most important role as the ontology serves as a starting edge structure for knowledge representation, and the procedure of ontology construction is one of the most critical research topics in the ontology processing.

V. Proposed Work

As context based searching is still not prevalent, so the main emphasize of our research is on that domain. Many popular search engines display all the information needed by the user without filtering anything. It also displays what is not required by the user and the result of any search goes up to lakhs. Our goal to make the user search more concise by displaying only information that is required by the user and discarding all that is irrelevant. So that the results displayed are focused results, i.e. only those information that is required by the user.

We divide our algorithm into three steps:

Step 1: In first step, we will construct ontology from the web repository.
Step 2: In second step, we will integrate this ontology with the semantic nets so that a focused document group can be created.
Step 3: In third step, we will accept the keywords to be searched and make search more concise by pruning the unwanted data and display the results based upon that along with its related context with the help of topic map that uses the ontology designed by us.
Our proposed work gives two fold benefit, firstly, only focused results are retrieved which reduces the number of results extracted and secondly, due to focused searching irrelevant results are pruned which reduces the time.

A. Ontology Construction

As the main objective of our research is to optimize the searching, by making changes in the way the user send his search keywords. Instead of searching in the whole web, our algorithm will search in the ontology built by us that is updated periodically. So before the actual web-searching starts, we should have a web-repository for the development of ontology (Structured knowledge about English word) in parallel.

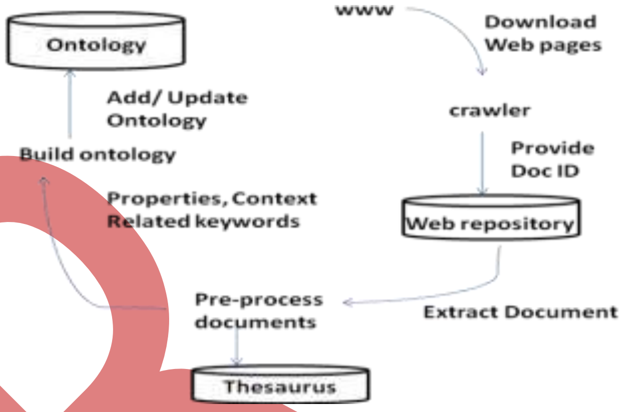

Figure 2:-Ontology construction

For building ontology we are using XML (Extensible Markup Language) which is a platform independent plain ASCII text file used as data description language. We have decided to use XML as it could be easily integrated with any of the web development language and it is very easy to use. To build a dynamic XML file, which could be automatically updated we have used C#.net language provided by Microsoft.

The advantage of this ontology is that once build, it could be used by any search engine to improve their performance in terms of results. Figure 1 Shows how the ontology commit, last step of ontology construction.

B. Building Topic Map

Based upon the keyword entered by the user, we will create a topic map using ontology build in step one. For doing so, we will again use C#.net that will retrieve the keyword along with its multiple contexts and its related topics. Thereafter displaying them on a web page in graphical form for making it easier for the user to extract what is desired by the user.

C. Pruning the Results

The main process on which our basic architecture relies to make the searching more focused and fast is pruning of the semantic network based on the ranking of context given by the user. Based on the ranking the network gets pruned displaying a specific topic map based result. For doing this we need a web repository from which result could be extracted, so we have used the web repository of Google, which is considered as the largest and fastest web repository. Using ASP.net we have customized the existing search technique to display more focused i.e. relevant results.

Algorithm for relevance calculation

- Repeat step 2 and 3 for each entity in sub entities table and for each document in the URL list.
- Calculate the number of times the entity Ei appears in each webpage (document) Dj .Let it be count.
- Store the count of each entity in each document in frequency table in database.
- Calculate total number of documents in URL list. Let it be N.
- For each document in the URL list repeat step 6-12.
- For each entity repeat steps 7-11.
- Calculate the number of documents that contain entity Ei. Let it be n
- Calculate the most repeated entity in document Dj. Let it be max.
- Read count of entity Ei in document Dj from frequency table in database.
- Calculate Ei entity weight in document Dj using the formula
- WEi = (count/max) * log (N/n)
- Store entity weight in entity document table in database.
- Calculate weight of document Dj by adding weight of all the entities in document Dj from Entitydocument table in database.

  WDj = WE1 + WE2 + . . . + WEm

Where m is the total number of entities in sub entity table. And store the weight of each document in Document weight table in database. If Document weight is greater than or equal to the threshold value then it is ranked and stored otherwise it is ignored.

End.

D. Proposed Architecture

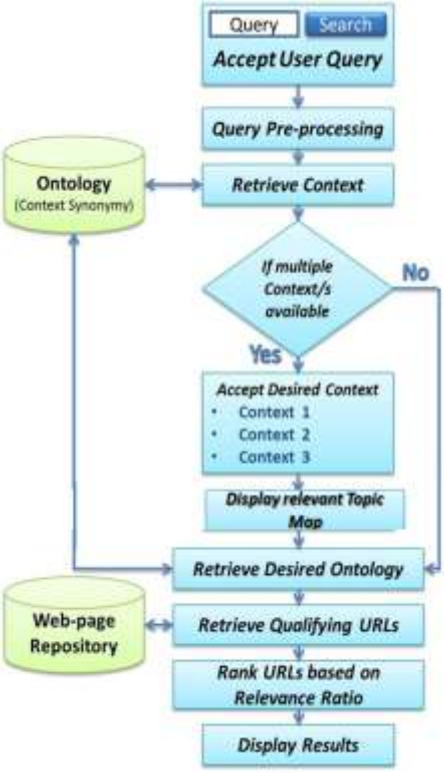

Figure shows the flow chart of our proposed architecture for focused crawling using ontology, which accept user query and after query preprocessing context is retrieved from ontology. If multiple contexts are available then accept desired context and display relevant topic map. From the relevant topic map we retrieve the desired ontology and on the basis of that ontology we retrieve the qualifying URLs form web page repository. Then we rank URLs based on relevance ratio and display the results.

VI. Conclusion

Our proposed model helps in the providing the solution to the most critical problem of information retrieval, Synonymy and Polysemy. This study proposes the systematic methodology to develop the ontology in a bottom-up style from engineering documents, called DocOnto (Document-based Ontology). Our methodology is mainly composed of three phases such as defining ontology, integrating the ontology with semantic networks and pruning the ontology for practically usage. This ontology can be updated and generalized using much easier process and is less time consuming and has specific definition of each word in the form of attributes.

It reduces the number of results extracted. Through focused searching irrelevant results are pruned which reduces the time. Displaying the multiple contexts and its related topic on a web page in graphical form, making it easier for the user to extract what is desired by the user. The advantage of our ontology is that once build, it could be used by any search engine. So it improved searching performance in terms of precision & relevance.

References:

[1] C. C. Aggarwal, F. Al-Garawi, and P. Yu., “Intelligent crawling on the World Wide Web with arbitrary predicates”, Hong Kong, in the proceedings of WWW10, 2001.

[2] D. Bergmark, C. Lagoze, and A. Sbityakov,”Focused crawls, tunneling, and digital libraries”, In ACM European Conference on Digital Libraries, Rome, 2002.

[3] Majidi Beseiso ,Abdul Rahim Ahmad and Roslan Ismail , “A New Architecture for Email Knowledge Extraction”, International Journal of web & Semantic Technology(IJWesT) Vol 3, No. 3 July 2012.

[4] J. Cho, H. Garc´_a-Molina, and L. Page,“Efficient crawling through URL ordering”, Computer Networks and ISDN Systems, vol 30(1-7), pp 161.172, 1998.

[5] M. Ester. M. Gross and H.-P. Kriegel, “Focused Web Crawling: A Generic Framework for specifying the User Interest and for Adaptive Crawling Strategies”, 27th International Conference on Very Large Databases, Roma, Italy, 2001.

[6] M. Diligenti,F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. “Focused Crawling using Context Graphs”, In VLDB-00, 2000.[7]”, In National.Symposium on Machine Learning (FGML '2000), Birlinghoven, 2000.

[7] Vagelis Hristidis, Yuhang Hu and Panagiotis G. Ipeirotis “Relevance-Based Retrieval Hidden-Web Text Databases without Ranking Support”. In IEEE

Transactions on Knowledge and Data Engineering, Vol 23, No. 10, pages 1555-1563, Oct 2011.

[8] A. Ardo, “Focused crawling in the alvis semantic search engine”. In ESWC ’05: Proceedings of the European Semantic Web Conference, pages 19–20, Heraklion, Greece, 2005.

[9] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: “A new approach to topic-specific web resource discovery”. In WWW ’99: Proceedings of the eighth international conference on World Wide Web, pages 1623–1640, New York, USA, 1999. Elsevier North-Holland, Inc.

[10] A. Gómez-Pérez, and D. Manzano-Macho, “A survey of ontology learning methods and techniques”,Deliverable 1.5, IST Project IST-2000-29243 - OntoWeb, 2003.

[11] T. Joachims, “Text categorization with support vector machines: learning with many relevant features”, Proceedings of the 10th ECML-1998, pp. 137–142.

[12] J. Rennie and A. McCallum “Using Reinforcement Learning to Spider the Web Efficiently”.,In International Conference on Machine Learning, ICML-99.

[13] A. Maedche and S. Staab, “Ontology Learning for the Semantic Web”, IEEE Intelligent Systems, Special Issue on the Semantic Web, March 2001,pp. 72 – 79.

[14] A. Maedche, G. Neumann and S. Staab, “Bootstrapping an Ontology-Based Information Extraction System”.,Studies in Fuzziness and Soft Computing, Intelligent exploration of the web, Springer, 2003, pp.345 – 359.

[15] B. Omelayenko, “Learning of ontologies for the Web: the analysis of existent approaches”, Proceedings of the international workshop on Web dynamics, London, 2001.

[16] M. Shamsfard and A.A. Barforoush, “The State of the Art in Ontology Learning”, The Knowledge Engineering Review, Cambridge Univ. Press, 2003, 18(4), pp. 293 – 316.

[17] S. T. Dumais, J. Platt, D. Heckerman and M. Sahami, “Inductive learning algorithms and representations for text categorization” , Proceedings of CIKM-98, pp 148–155.

[18] M. Ehrig and A. Maedche, “Ontology-focused crawling of web documents”, In SAC ’03: Proceedings of the 2003 ACM symposium on applied computing, pages 1174–1178, New York, NY, USA, 2003.

[19] S. M. Pahlevi and H. Kitagawa.”, Taxonomy-based adaptive web search method”, In ITCC ’02: Proceedings of the International Conference on Information Technology: Coding and Computing, page 320, Washington, DC, USA, 2002, IEEE Computer Society.

[20] J. Qin, Y. Zhou, and M. Chau, “Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method”, In JCDL ’04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, pages 135–141, New York, NY, USA, 2004.

[21] J.Graupmann, M. Biwer, C. Zimmer, P. Zimmer, M. Bender, M. Theobald and G. eikum. Compass, “A concept-based web search engine for html, xml, and deep web data”, In VLDB ’04: Proceedings of the Thirtieth international conference on Very large data bases, pages 1313–1316.VLDB Endowment, 2004.

[22] J. M. Park, J. H. Nam, Q. P. Hu, H. W. Suh, “Product Ontology Construction from Engineering Documents”, IEEE, 2008.

[23] Thomas R. Gruber, “A translation approach to portable ontology specifications, Knowledge Acquisition” Vol 5 , no. 2, pp 199–220, 1993.

[24] M. Ushold and M. Gruninger, “Ontologies: Principles, methods and applications”, The Knowledge Engineering Review, 1996.

[25] Spyns P. and Meersman R., “A survey on ontology alignment and merging”, OntoBasis Deliverable #D1.5, STAR Lab, Brussel, 2003.

[26] Nicolas Martin, Khaled Khelif, IPCC, CASSIDIAN, Focused Crawling using name disambiguation on search engine results, 2011 Europian Intelligence and Security Informatics Conference, 978-0-7695-4406-9/11, pp 340-345, IEEE-2011.

[27] Ioannis Avraam, Ioannis Anagnostopoulos, Aristotle University of Thessaloniki, Greece, A Comparision over focused web crawling strategies, 2011-Panhellenic Conference on Informatics, 978-0-7695-4389-5/11, pp 245-249, IEEE-2011.

[28] Daniel Osuna-Ontiveros, Ivan Lopez-Arevalo, Victor Sosa, A Semantic Information Retreival model for focused crawling, 2011-7th International Conference on next generation web services practice, 978-1-4577-1127-5/11, pp 285-289, IEEE - 2011.

[29] Hiep Phuc Luong, Susan Gauch, Qiang Wang, University of Arkansas, Ontology Based Focused Crawling, International Conference on Information, Process & Knowledge Management, 978-0-7695-3531-9/09, pp 123-128, IEEE-2009.

[30] Hong-Wei Hao, “An improved topic relevance algorithm for focused crawling”, Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference, 2011

[31] Mohsen, Jamali, Hassan Sayyadi, Babak Bagheri Hariri, Hassan Abolhassani, “A method for focused crawling using combination of link structure and content similarity”, International conference on Web-Intelligence, 0-7695-2747-7/06, IEEE-2006.

[32] Mehdi Ravakhah, Mohsan Kamyar, "Semantic similarity based focused crawling", 2009 First International Conference on Computational Intelligence, Communication Systems and Networks, 978-0-7695-3743-6/09, pp 448-452, IEEE-2009.

[33] Hong-Wei Hao, Cui-Xia Mu, Xu-Cheng Yin, Shen Li, Zhi-Bin Wang, "An improved topic relevance algorithm for focused crawling", 978-1-4577-0653-0/11, pp 850-855, IEEE-2011.

[34] Devashis Hati, Biswajit Sahoo, Amritesh Kumar, "Adaptive focused crawling based on link analysis", 2010 Second International Conference on Education Technology & Computer (ICETC), 978-1-4244-6370-1/10, pp 455-459, IEEE-2010.