ISSN: 2321-7782 (Online)

Volume 2, Issue 12, December 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Ontology Based semantic web Crawler Mechanism for Information Discovery

Swapnil V. Patil[1]
Dept. of Computer engineering
JSPM'S JSCOE
Pune – India

Sharmila M. Shinde[1]
Dept. of Computer engineering
JSPM'S JSCOE
Pune – India

Abstract: Due to availability of abundant data on web, searching has a significant impact. Ongoing researches place emphasis on the relevancy and robustness of the data found, as the discovered patterns proximity is far from the explored. In spite of their relevance pages for any search topic, the results are huge to be explored. Also the users' perspective differs from time to time from topic to topic. Usually ones' want is others unnecessary. Crawling algorithms are thus crucial in selecting the pages that satisfies the users' needs. This paper reviews the researches on web crawling algorithms used on searching.

Keywords: web crawling algorithms, crawling algorithm survey, search algorithm, ontology learning, hybrid crawler.

I. Introduction

Text mining mostly used to find out unknown information from natural language processing and data mining by applying various techniques. In this technique for discovering the importance of term in document, term frequency of term is calculated. Sometime we can notice that two terms having same frequency in document but one term leads more meaning than other, for this concept based mining model is intended. In proposed model three measures are evaluated for analyzing concept in sentence, document and corpus levels. Semantic role labeler is mostly associated with semantic terms. The term which has more semantic role in sentence, it's known as Concept. And that Concept may be either word or phrase depending on sentence of semantic structure. When we put new document in system, the proposed model discover concept match by scanning all new documents and take out matching concept. The similarity measures that are used for concept analysis on sentence, document and corpus level exceeds similarity measures depending on the term analysis model of document. The results are measured by using F-measure and Entropy. We will use this model for web documents.

II. Literature survey

In this section, we briefly introduce the fields of semantic focused crawling and ontology-learning-based focused crawling, and review previous work on ontology learning-based focused crawling. A semantic focused crawler is a software agent that is able to traverse the Web, and retrieve as well as download related Web information on specific topics by means of semantic technologies [6], [7][15]. Since semantic technologies provide shared knowledge for enhancing the interoperability between heterogeneous components, semantic technologies have been broadly applied in the field of industrial automation [8]–[10]. The goal of semantic focused crawlers is to precisely and efficiently retrieve and download relevant web information by automatically understanding the semantics underlying the Web information and the semantics underlying the predefined topics.

A survey conducted by Dong *et al.* [11] found that most of the crawlers in this domain make use of ontologies to represent the knowledge underlying topics and Web documents. However, the limitation of the ontology-based semantic focused crawlers is that the crawling performance crucially depends on the quality of ontologies. Furthermore, the quality of ontologies may be affected by two issues. The first issue is that, as it is well known that ontology is the formal representation of specific domain knowledge [12] and ontologies are designed by domain experts, a discrepancy may exist between the domain experts' understanding of the domain knowledge and the domain knowledge that exists in the real world. The second issue is that knowledge is dynamic and is constantly evolving, compared with relatively static ontologies. These two contradictory situations could lead to the problem that ontologies sometimes cannot precisely represent real-world knowledge, considering the issues of differentiation and dynamism. The reflection of this problem in the field of semantic focused crawling is that the ontologies used by semantic focused crawlers cannot precisely represent the knowledge revealed in Web information, since Web information is mostly created or updated by human users with different knowledge understandings, and human users are efficient learners of new knowledge. The eventual consequence of this problem is reflected in the gradually descending curves in the performance of semantic focused crawlers.

In order to solve the defects in ontologies and maintain or enhance the performance of semantic-focused crawlers, researchers have begun to pay attention to enhancing semantic- focused crawling technologies by integrating them with ontology learning technologies. The goal of ontology learning is to semi-automatically extract facts or patterns from a corpus of data and turn these into machine-readable ontologies [13].

Various techniques have been designed for ontology learning, such as statistics-based techniques, linguistics (or natural language processing)-based techniques, logic-based techniques, etc. These techniques can also be classified into supervised techniques, semi-supervised techniques, and unsupervised techniques from the perspective of learning control. Obviously, ontology-learning-based techniques can be used to solve the issue of semantic-focused crawling, by learning new knowledge from crawled documents and integrating the new knowledge with ontologies in order to constantly refine the ontologies.

III. WEB CARWLER STRATEGIES

1. Breadth First Search Algorithm:

This algorithm aims in the uniform search across the neighbour nodes. It starts at the root node and searches the all the neighbour nodes at the same level. If the objective is reached, then it is reported as success and the search is terminated. If it is not, it proceeds down to the next level sweeping the computation time. bEMADS and gEMADS these two algorithms are used based on Gaussian mixture model. Both resumes data into sub cluster and after that generate Gaussian mixture. These two algorithms run several orders of magnitude faster than maximum with little loss of search across the neighbour nodes at that level and so on until the objective is reached. When all the nodes are searched, but the objective is not met then it is reported as failure.

Breadth first is well suited for situations where the objective is found on the shallower parts in a deeper tree. It will not perform so well when the branches are so many in a game tree, especially like chess game and also when all the path leads to the same objective with the same length of the path [2][3].

Andy yoo et al [4] proposed a distributed BFS for numerous branches using Poisson random graphs and achieved high scalability through a set of clever memory and communication optimizations.

2. Depth First Search Algorithm

This powerful technique is systematically traversing through the search by starting at the root node and traverse deeper through the child node. If there are more than one child, then priority is given to the left most child and traverse deep until no more child is available. It is backtracked to the next unvisited node and then continues in a similar manner [5].

This algorithm makes sure that all the edges are visited once breadth [6]. It is well suited for search problems, but when the branches are large then this algorithm takes might end up in an infinite loop [3].

3. Page Rank Algorithm

Page rank algorithm determines the importance of the web pages by counting citations or back links to a given page [7]. The page rank of a given page is calculated as

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... +

PR(Tn)/C(Tn))

PR(A) Page Rank of a Website,

d damping factor

T1,….Tn links

Yongbin Qin and Daoyun Xu [8] proposed an algorithm, taking the human factor into consideration, to introduce page belief recommendation mechanism and brought forward a balanced rank algorithm based on PageRank and Page belief recommendation which ultimately attaches importance into the subjective needs of the users; so that it can effectively avoid topic drift problems. Tian Chong [9] proposed a new type of algorithm of page ranking by combining classified tree with static algorithm of PageRank, which enables the classified tree to be constructed according to a large number of users' similar searching results, and can obviously reduce the problem of Theme-Drift, caused by using PageRank only, and problem of outdated web pages and increase the efficiency and effectiveness of search. J.Kleinberg [10] proposed a dynamic page ranking algorithm. Shaojie Qiao [11] proposed a new page rank algorithm based on similarity measure from the vector space model, called SimRank, to score web pages. They proposed a new similarity measure to compute the similarity of pages and apply it to partition a web database into several web social networks (WSNs).

4. Genetic Algorithm

Genetic algorithm is based on biological evolution whereby the fittest offspring is obtained by crossing over of the selection of some best individuals in the population by means of fitness function. In a search algorithm solutions to the problem exists but the technique is to find the best solution within specified time [12]. [13] Shows the genetic algorithm is best suited when the user has literally no or less time to spend in searching a huge database and also very efficient in multimedia results. While almost all conventional methods search from a single point, Genetic Algorithms always operates on a whole population. This contributes much to the robustness of genetic algorithms. It reduces the risk of becoming trapped in a local stationary point [14].The applicability of Genetic Algorithms by various researchers [16], [17], [18], [19] has been depicted in [20].

5. Naive Bayes classification Algorithm

Naive Bayes algorithm is based on Probabilistic learning and classification. It assumes that one feature is independent of another [21]. This algorithm proved to be efficient over many other approaches [22] although its simple assumption is not much applicable in realistic cases [21]. Wenxian Wang et al [23] proposed an efficient crawler based on Naive Bayes to gather many relevant pages for hierarchical website layouts. Peter Flach and Nicolas Lachiche [24] presented Naive Bayes classification of structured data on artificially generated data.

6. HITS Algorithm

This algorithm put forward by Kleinberg is previous to Page rank algorithms which uses scores to calculate the relevance [25]. This method retrieves a set of results for a search and calculates the authority and hub score within that set of results. Because of these reasons this method is not often used [1].

IV. CONCEPTS HELPFUL TO MEASURE THE SIMILARITIES

The techniques prescribed in the previous work are used for document clustering. But it is only for documents present on system. In the proposed system we are going to use web documents and we will get the clustered output and that have shown in fig.1. This concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. A web document is given as the input to the proposed model. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based on the Prop Bank notations. After running the semantic role labeler, labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document, and corpus level. In the concept-based mining model, a labeled terms either word or phrase is considered as concept.

The proposed model contains the following modules

A. Web Document

Web document is given as Input to the given system. Here user can give any query to the browser. Pure HTML pages are selected by removing extra scripting. Web pages contain data such as hyperlinks, images, script. So it is necessary to remove such unwanted script if any, during the time when a page is selected for processing. The HTML code is then transferred into XML code. On that document next process is processed that is Text pre-processing or data processing.

B. Data Processing

First step is separate sentences from the documents. After this label the terms with the help of Prop Bank Notation. With the help of Porter algorithm remove the stem word and stop words from the terms.

C. Concept Based Analysis

This is important module of the proposed system. Here we have to calculate the frequencies of the terms.

Conceptual term frequency (ctf), Term frequency (tf) and Document frequency (df) are calculated. The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than document only.

1. Sentence based concept analysis

For analyzing every concept at sentence level, concept based frequency measure; called conceptual term frequency is used.

1.1 Calculating ctf in sentence s

Ctf is the number of occurrences of concept c in verb structure of sentence s. If concept c frequently appears in structure of sentence s then it has principal role of s.

1.2 Calculating ctf in document d

A concept c can have many ctf values in different sentences in the same document d. Thus, the ctf value of concept c in document d is calculated by:

$$ctf = \frac{\sum_{n=1}^{sn} ctf_n}{sn},$$

2. Document based concept analysis

   For analyzing concepts at document level term frequency tf in original document is calculated. The tf is a local measure on the document level.

3. Corpus based concept analysis

   To calculate concepts from documents, document frequency df is used. Document Frequency df is the global measure. With the help of Concept based Analysis Algorithm we can calculate ctf, tf, df.

D. Similarity Approach

This module mainly contains three parts. Concept based similarity, Singular Value Decomposition and combined based similarity it contains. Here we get that how many percentage of concept math with the given web document.

E. Concept Based Similarity

A concept-based similarity measure depends on matching concept at sentence, document, and corpus instead of individual terms. This similarity measure is based on three main aspects. First are analyzed label terms that capture semantic structure of each sentence. Second is concept frequency that is used to measure participation of concept in sentence as well as document. Last is the concepts measured from number of documents. Concept based similarity between two document is calculated by:

$$sim_c(d_1, d_2) = \sum_{i=1}^{m} max\left(\frac{l_{i_1}}{Lv_{i_1}}, \frac{l_{i_2}}{Lv_{i_2}}\right) \times weight_{i_1} \times weight_{i_2}$$

Term frequency is calculated by following formula:

$$tf\ weight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tf_{ij})^2}},$$

F. Clustering Techniques

This module used three main basic techniques like Single pass, Hierarchical Agglomerative Clustering, and K-Nearest Neighbor. With the help of these techniques we can get that which cluster is having highest priority.

G. Output Cluster

Last module is the output Cluster. After applying the clustering techniques we get clustered document. That will help to find out main concepts from the web document.

V. SYSTEM IMPLEMENTATION

The proposed system model illustrates flow of implementation. First, web document given input to the system where, HTML pages are collected and their XML conversion is carried out. In the second module that is in Text Processing carried out separate sentences, label terms, and removing stop words and stem words. Third module Concept based analysis measures conceptual term frequency (ctf), term frequency (tf), and document frequency (df). Next module concept based document similarity find out how many percentage of concept is similar to the given concept.

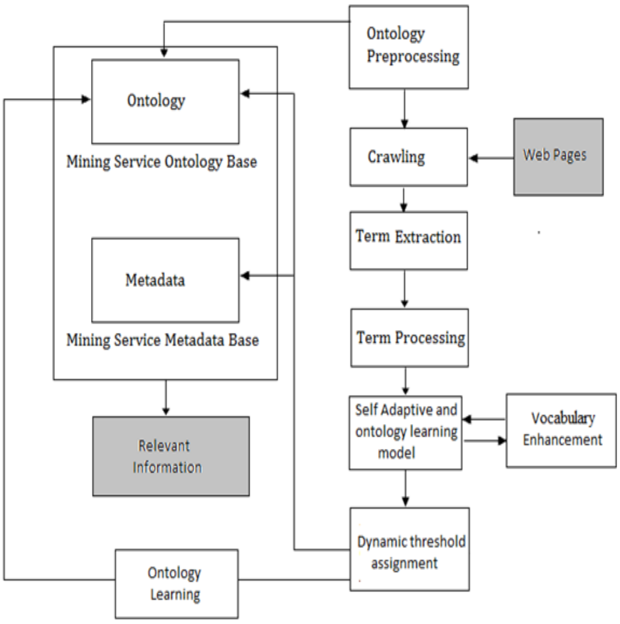

Fig. 1 General Block Diagram

VI. Conclusion

The main objective of the review paper was to throw some light on the web crawling algorithms. We also discussed the various search algorithms and the researches related to respective algorithms and their strengths and weaknesses associated. We believe that all of the algorithms surveyed in this paper are effective for web search, but the advantages favours more for Genetic Algorithm due to its iterative selection from the population to produce relevant results.

Acknowledgement

I express my sincere thanks to Sharmila M. Shinde whose supervision, inspiration and valuable guidance helped me a lot to complete my review paper. Her guidance proved to be the most valuable to overcome all the hurdles in the fulfilment of this paper.

References

1. Alessio Signorini, "A Survey of Ranking Algorithms" retrieved from http://www.divms.uiowa.edu/~asignori/phd/report/asurvey-of-ranking-algorithms.pdf 29/9/2011
2. Steven S. Skiena "The Algorithm design Manual" Second Edition, Springer Verlag London Limited, 2008, Pg 162.
3. Ben Coppin "Artificial Intelligence illuminated" Jones and Barlett Publishers, 2004, Pg 77.
4. Andy Yoo,Edmond Chow, Keith Henderson, William McLendon, Bruce Hendrickson, ÄUmit CatalyÄurek "A Scalable Distributed Parallel Breadth-First Search Algorithm on BlueGene/L" ACM 2005.
5. Alexander Shen "Algorithms and Programming: Problems and solutions" Second edition Springer 2010, Pg 135
6. Narasingh Deo "Graph theory with applications to engineering and computer science" PHI, 2004 Pg 301
7. Sergey Brin and Lawrence Page "Anatomy of a Large scale Hypertextual Web Search Engine" Proc. WWW conference 2004
8. Yongbin Qin and Daoyun Xu "A Balanced Rank Algorithm Based on PageRank and Page Belief recommendation"
9. TIAN Chong "A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine" Proc International Conference on Computer Application and System Modeling (ICCASM 2010)
10. J.Kleinberg "Authoritative sources in a hyperlinked environment", Proc 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
11. Shaojie Qiao, Tianrui Li, Hong Li and Yan Zhu, Jing Peng, Jiangtao Qiu "SimRank: A Page Rank Approach based on similarity measure" 2010 IEEE
12. S. N. Sivanandam, S. N. Deepa "Introduction to Genetic Algorithms" Springer, 2008, pg 20
13. S.N. Palod, Dr S.K.Shrivastav,Dr P.K.Purohit "Review of Genetic Algorithm based face recognition" International Journal of Engineering Science and Technology (IJEST) Vol. 3 No. 2 Feb 2011
14. Deep Malya Mukhopadhyay, Maricel O. Balitanas, Alisherov Farkhod A.,Seung-Hwan Jeon, and Debnath Bhattacharyya "Genetic Algorithm: A Tutorial Review" International Journal of of Grid and Distributed Computing Vol.2, No.3, September, 2009.
15. Hai Dong, Member, IEEE, and Farookh Khadeer Hussain" Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery" IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 10, NO. 2, MAY 2014.

16. Zacharis Z. Nick and Panayiotopoulos Themis, Web Search Using a Genetic Algorithm, IEEE Internet computing,1089-7801/01c2001, 18-25, IEEE

17. Ramakrishna Varadarajan, Vagelis Hristidis, and Tao Li , Beyond Single-PageWeb Search Results,IEEE Transactions on knowledge and data engineering, 20(3) ,411 - 424, 2008

18. Judit BarIlan, Comparing rankings of search results on the Web, Information Processing and Management 41 (2005) 1511–1519

19. Adriano Veloso, Humberto M. Almeida, Marcos Goncalves, Wagner Meira Jr.,Learning to Rank at Query- Time using Association Rules, SIGIR'08, 267-273 , 2008, Singapore.

20. S.Siva Sathya and Philomina Simon," Review on Applicability of Genetic Algorithm to Web Search" International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October2009

21. Harry Zhang "The Optimality of Naive Bayes" American Association for Artificial Intelligence 2004. Rich Caruana, Alexandru Niculescu-Mizil "An Empirical Comparison of Supervised Learning Algorithms" Proc 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

22. Wenxian Wang, Xingshu Chen, Yongbin Zou, Haizhou Wang, Zongkun Dai "A Focused Crawler Based on Naive Bayes Classifier" Third International Symposium on Intelligent Information Technology and Security Informatics, 2010

23. Peter A. Flach and Nicolas Lachiche "Naïve Bayesian Classification of Structured Data" Machine Learning, Kluwer Academic Publishers

24. Kleinberg, John "Hubs, Authorities, and Communities" ACM computing survey,1998.

AUTHOR(S) PROFILE

Swapnil V. Patil received the B.E. degree in Computer Science Engineering from Bharati Vidyapeeth in 2010. Currently he is pursuing M.E. in Computer Engineering in JSPM'S JSCOE, Pune.

Sharmila M. Shinde, received the M.E. degree in Computer Engineering from Bharati Vidyapeeth in 2004. Currently she is an Associate professor and Head of Department of Computer Engineering in JSPM'S JSCOE, Pune.