A SURVEY ON FOCUSED CRAWLER FOR SEED URL SELECTION

Mrs. S. Nithyapriya
PG Scholar,
Computer Science and Engineering,
SNS College of Technology,
Coimbatore, Tamilnadu, India

Dr. T. Kalaikumaran
Professor and Head,
Computer Science and Engineering,
SNS College of Technology,
Coimbatore, Tamilnadu, India

Dr. S. Karthik
Professor and Dean,
Computer Science and Engineering,
SNS College of Technology,
Coimbatore, Tamilnadu, India

*Abstract***—** ***The World Wide Web (WWW) contains large amount of information due to the increase of web pages. Extracting relevant information from web requires suitable search strategies. Conventional search engine are not suitable anymore. Web documents that are suitable to predefined user interest topics are extracted from web by using focused crawler based on seed URLs selection. Ontology is a formal specification of knowledge by a set of concepts with in a domain and their semantic relationship. Different Ontologies are developed by developer for a same domain differently. In this paper we survey and review issues related to focused crawling strategies in semantic web.***

*Keywords***—** ***Focused Crawler, Seed URLs, Ontology, Semantic web.***

I. INTRODUCTION

Semantic Web: The semantic web describes the meaning of information that can be understood by people and computer machine. The semantic web uses RDF to describe web resources with background in logic and artificial intelligences. The utility of semantic web depends on three issues such as existence of data; user can retrieve the data, quality. Ontology is a formal explicit representation of concepts in domain properties of each concept describes the characteristics and attributes of the concepts known as slots. Conceptualization is a description of concepts and relationship exists. Ontology is the platform for sharing the knowledge of domain that helps the machine to make intelligent decision.

Limitations of Current Web:

- The web content lacks proper structure for representation.
- Uncertainty of information.
- Usability to deal with enormous number of users and content ensuring trust at all levels.
- Incapability of machines to understand the information due to lack of a universal format.

Focused Crawler: Web crawler is an application program download web pages when seed URL is given as input. Focused crawler crawls through domain specific pages. Focused crawlers selectively look for large number of relevant documents on that specific domain and effectively discard irrelevant documents and hence leading to significant savings in both computation and communication resources, and high quality retrieval results [15]. Focused crawling approach improves the precision and recall of expert search on the Web. Early focused crawlers use domain keywords to find the page is relevant or not. The focused crawler is enhanced by using ontology to detect the relevant score for links before downloading. They extract and order the previously download documents using ontology by computing the page relevance.

Challenges in Focused Crawler:

- Crawler wants to crawl through large number of dynamic web pages.
- Crawler have to maintain count how frequently revisit page.
- It should select correct URL for extracting relevant information.
- Rank and order the relevant URLs to determine the relevance of a web page.

Seed URLs: Seed URL with topic of user queries are the main parameter of a focused web crawler. We start with web page to find link to other pages. Starting URL called Seed URL. Selecting proper seeds increase number of pages, so crawler will discover and result in collection with more good pages and less bad pages. Unvisited URL called frontier. Select seed URL which has high hub score and authority score. Seed URL connect many other web pages to assure higher recall and precision of focused crawler.

The seed URLs of an interested topic can be deduced from a search result of a selected search engine. Seed URLs do not contain relevant web pages; the crawler has less possibility to find many other relevant ones.

This paper is organized as follows. Some related research works are briefly reviewed in Section 2. Conclusion is drawn in section 3.

II. RELATED WORK

This section deals with the issues of selecting seed URL of Focused Crawler. The Seed URL selection is one of the bottlenecks in the research of semantic field. Recently, some interesting techniques and methodologies are focus on the interoperability among the domain specific data sources.

2.1 HITS

Jon Kleinberg proposed the HITS algorithm. HITS stands for Hyperlink-Induced Topic Search. HITS are a link analysis algorithm for rating web pages. Good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs [3]. Selecting seed URL set based on authority web pages. [12] In the HITS algorithm, the most relevant pages are retrieved to the search query called the root set which is obtained by taking the top n pages returned by a text-based search algorithm. A base set is formed by augmenting the root set with all the web pages that are linked from it. The web pages in the base set and all hyperlinks form a focused sub graph. The HITS computation is performed only on this focused sub graph. Rank is calculated by computing hub and authorities score of the pages in order of their relevance. Authority score estimates the value of the content of the page. Hub score estimates the value of its links to other pages.

Advantages

- Returned pages have high relevancy and importance.
- Ability to rank pages according to the query topic, able to provide more authority and hub pages.

Disadvantages

- Query dependent
- Problem of topic drift
- Time consuming
- Spammed easily

2.2 Formal Concept Analysis

Wille's FCA is a conceptual framework and based on lattice theory [2] .The FCA method defines formal contexts to represent the relationships between objects and attributes in a domain. FCA analyzes data which describe relationship between a particular set of objects and attributes. The applications of FCA method are text processing, ontology generation, ontology merging etc. It includes contributing context, and translating context into the concept lattice. Formal context in FCA is a triple K = (G, M, I) where G is a set of objects, M is a set of attributes and the binary relation I ⊆ G × M shows which objects possess which attributes. A concept lattice is a collection of formal concepts in the data which are hierarchically ordered by a sub concept-super concept relation.

Advantages

- Helps to discover meaningful data in binary relations.
- Can be visualized with Concept Lattices.

Disadvantages

- Context with fewer attributes.
- Extraction of all concepts infeasible for large contexts.

2.3 HCONE-merge

Kotis and Vouros [10] proposed HCONE-merge approach which is used for ontology merging. This approach uncovers the intended informal meaning of concepts specified in ontology by mapping them to WordNet senses. WordNet senses realize the informal, human-oriented intended meaning of the corresponding concepts. Linguistic and structural knowledge about ontologies are exploited by the Latent Semantics Indexing method (LSI).The HCONE-merge approach requires humans to validate the computed intended meaning of every concept in the ontology.

Advantage

- It supports the mapping/merging of ontologies in absence of reference ontology.

Disadvantage

- Highly technical domain ontologies terms do not have an entry in WordNet resulting in poor performance of the method.

2.4 Shark-Search

In Fish-Search algorithm, Web agents are like the fishes in sea. If the page is not relevant, its child links receive a low preferential value. Fish gain energy when a relevant document found. But the limitations in Fish-Search are it assigns discrete relevance scores and low discrimination of the priority of pages. Shark-Search is a modification of Fish-search by assigning real-valued relevance scores based on ancestral relevance score, anchor text and textual context of the link. Child inherits a discounted value of the score of its parent, and this score is combined with a value based on the anchor text.

2.5 PageRank

The original PageRank algorithm was proposed by Lawrence Page and Sergey Brin [16]. PageRank depends on link structure of the web pages. PageRank does not rank web sites as a whole, but is determined for each page individually. Further, the PageRank of page A is recursively defined by the PageRank of those pages which link to page A. The Page Rank considers the back link in deciding the rank score. The page rank is given by

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (1)$$

PR(A) is the PageRank of page A,

PR(Ti) is the PageRank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti

d is a damping factor which can be set between 0 and 1.

Advantages

- Query-time cost of incorporating precomputed PageRank importance score for a page is low.
- Less susceptible to localized link spam.

Disadvantages

- Results come at the time of indexing and not at the query time.
- Dangling Links.

2.6 InfoSpider

Monge[6] proposed a web crawler-InfoSpider, dynamic web search multi agent system. The user queries are submitted to a general search engine, InfoSpider returned result set as the candidate set of the seed URLs. InfoSpider complement traditional index based search engines using agents at the user side. An agent is initialized for every link and analyses the links by computing the similarity of the text around the link with the query using neural net. The next link is chosen with a probability proportional to the similarity score. The neural net weights are adjusted by the relevance of the new page's content so that the agent updates its knowledge.

2.7 Breadth-First

Breadth-First algorithm uses the frontier as a FIFO queue and maintains the crawling links in the order in which they are encountered. It maintains a frontier of known URLs as a priority queue sorted by the cosine similarity between the topic and the page where the URL was found. The crawler can add only one link from a crawled page when the frontier is full. Breadth-first crawling checks each link on a page before proceeding to the next page.

2.8 Decision Tree

Decision tree is the classification technique. This classifier is used for computing the relevance of the pages in the graph. It first creates a decision tree based on training data set. The decision tree construction is done by the ID3 method, analyzing the terms of the anchor text of all the links in the graph. The link prioritization is done with the help of a decision tree which is trained by a web graph provided by the user. It is not suitable for large scale focused crawling; instead, it is designed to crawl into a limited portion of the web.

Disadvantage

- Decision trees create a complex model which cannot be generalized well and to overcome this we need to implement pruning.

2.9 Eigen Rumor Algorithm

The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, an Eigen Rumor algorithm [11] provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of Eigen vector.

Advantage

- High for blog ranking.

Disadvantage

- It is most specifically used for blog ranking not for web page ranking.

III. CONCLUSION

Selecting suitable Seed URL for Focused Crawler is an important issue in semantic web. This paper presented a scope of various Focused crawling strategies strength and weaknesses. Future enhancements are made to overcome challenges of existing approaches competently. User-interest Ontology based semantic search engine are able to handle the web search problems successfully.

References

[1] Ahmed Patel, Nikita Schmidt, "Application of structured document parsing to focused web crawling", ELSEVIER 33 (2011) 325-331.

[2] B.Ganter,R.Wille, Formal Concept Analysis: Mathematical Foundations,Springer-Verlag, Berlin, 1999.

[3] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze (2008). "Introduction to Information Retrieval". Cambridge University Press. Retrieved 2008-11-09.

[4] David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas and Yannis Avrithis, "Personalized Content Retrieval in Context Using Ontological Knowledge", IEEE transactions on circuits and systems for video technology,(2007) vol. 17, no. 3.

[5] Filippo menczer, gautam pant, padmini srinivasan ,"Topical Web Crawlers: Evaluating Adaptive Algorithms", ACM Transactions on Internet Technology, (2004) 378–419.

[6] F.Menczer,A.E. Monge, Scalable web search by adaptive online agents, an Infospider case study, in: The Proceeding of Intelligent Information Agents, Agent-Based Information Discovery and Management on the Internet, Springer,Berlin, 1999,pp.323–347.

[7] Geir Solskinnsbakk, Jon Atle Gulla" Combining ontological profiles with context in information retrieval",ELSEVIER,69(2010) 251-260.

[8] Hai Dong,Farookh Khadeer Hussain,"Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery“ IEEE Transactions On Industrial Informatics, VOL. 10, NO. 2, MAY (2014).

[9] Hele-Mai Haav,“A Semi-automatic Method to Ontology Design by Using FCA”, VˇSB – Technical University of Ostrava(2004) 13-24.

[10] K. Kotis, G. Vouros, K. Stergiou, Towards automatic merging of domain ontologies: the HCONE-merge approach? International Journal of Web Semantics 4(1) (2006) 60–79.

[11] Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, “The EigenRumor Algorithm for Ranking Blogs”, In WWW 2005 Annual Workshop on the Web logging Ecosystem, 2005.

[12] Kleinberg, Jon (1999). "Authoritative sources in a hyperlinked environment" . Journal of the ACM 46 (5): 604–632

[13] Marek Obitko, Vaclav Snasel, and Jan Smid ,”Ontology Design with Formal concept Analysis”, VSB – Technical University of Ostrava(2004) 111-119.

[14] M. Ehrig, A. Maedche, "Ontology-focused crawling of Web documents", Proceedings of the 2003 ACM SymposiumComputing, pp. 1174-1178, USA, (2003).

[15] M. Kumar and R. Vig, “Design of CORE: context ontology rule enhanced focused web crawler”, International Conference on Advances in Computing, Communication and Control (ICAC3□09) pp. 494-497, 2009.

[16] Page, L., Brin, S., Motwani, R. & Winograd, T., 1998. The PageRank Citation Ranking: Bringing Order to theWeb. Stanford Digital Library Technologies Project.

[17] Q. Cheng, W. Beizhan and W. Pianpian. "Efficient focused crawling strategy using combination of link structure and content similarity",IEEE (2008).

[18] Rung-Ching Chen, Cho-Tscan Bau, Chun-Ju Yeh,"Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques", ELSEVIER,11(2011) 1908-1923.

[19] Xiangping Kang , Deyu Li , Suge Wang ,”Research on domain ontology in different granulations based on concept lattice”, Elsevier B.V.(2012) 152-161.

[20] YaJunDu,YuFengHai,ChunZhiXie,XiaoMingWang, “An approach for selecting seed URLs of focused crawler based on user-interest ontology”, ELSEVIER 14 (2014) 663-676.

[21] YaJun Dua,*, HaiMing Li ,“Strategy for mining association rules for web pages based on formal concept analysis” Elsevier B.V.(2009) 772–783.