Ontology Based Web Crawler for Specific Domain

[1]Swasti Singhal, [2]Neeraj Kumar
[1,2]Dept. of Information Technology, Galgotias College of Engineering and Technology

Abstract
As available of large amount of data in World Wide Web (www), browse of any specific domain related topic takes too much time and also contain irrelevant web pages which are not undesirable. For crawler it is not easy task to download only data mining related web pages. Ontology is the technique to access only data mining related web pages or domain specific pages. So the basic goal of "ontology based web crawler for domain specific" is to select and seek out the web pages that fulfill user's requirement for example data mining related web pages. The link analysis algorithms like page ranking algorithm and other metrics are use to prioritize the URLs based on their ranking and selection policies for downloading most specific web pages.

Keywords
Ontology, Web Crawler, Page Ranking, World Wide Web (www), Best First Search

I. Introduction

World Wide Web is collection of web document. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks. There are thousands of pages available over the web. To fetch or to get the desired information there must be some particular way or mechanism. This process to look for desired documents or information available all over the globe is called web searching. We need a specified application or tool that can easily tell us that the document we are searching for is available over the internet or not. Such applications are called search engines.

Search Engines are special programs that are implemented over the servers and help users all over the globe to search for any kind of data available over the World Wide Web. Search engines comprises of various components that do their respective work in order to return the relevant documents available over the web corresponding to the users query. The most crucial components that a search engine contains are Web Crawler, Indexer, Ranker, and User Interface.

A Web crawler [1] is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, Web spiders, Web robots. This process is called Web crawling or spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for sending spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the "crawl frontier". URLs from the frontier are recursively visited according to a set of policies.

Indexer indexes all the web pages and documents crawled by web crawler and store this relevant data in the form of table or related type. Indexer indexes and maintains the information available over the web and also stores what kind of data is available on the indexed page. Indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, physics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is web indexing.

A Page Rank results from a mathematical algorithm based on the graph, the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high Page Rank receives a high rank itself. If there are no links to a web page there is no support for that page.

Numerous academic papers concerning Page Rank have been published since Page and Brin's original paper [2]. This paper, the Page Rank concept has proven to be vulnerable to manipulation, and extensive research has been devoted to identifying falsely inflated Page Rank and ways to ignore links from documents with falsely inflated Page Rank.

While ranking the web pages special algorithm is implemented, this varies from one search engine to other search engine. Google uses the famous "Page Ranking algorithm". Generally search engine ranks the pages based on the information gather through their hyperlinks. More the number of hyperlinks more rank a page holds.

A search engine is a document retrieval system which helps find information stored in a computer system, such as in the World Wide Web (WWW), inside a corporate or proprietary network, or in a personal computer. Surveys indicate that almost 25% of Web searchers are unable to find useful results in the first set of URLs that are returned. The term ontology is an old term used in the field of Knowledge Representation, Information Modelling, etc. Typically ontology is a hierarchical data structure containing relevant entities, relationships and rules within a specific domain. Tom R. Gruber [3] defines ontology as a specification of a conceptualization.

II. Literature Review

During searching the data, a lot of results appear. User can not visit all the web results and they no much time for analysis. So search engine have big task to sort out the web pages according to the user interest.

Retrieving effective content from the web is a crucial task. Users often look at only a few top hits, making the precision achieved by the ranking algorithm. Early search ranked pages principally based on their lexical similarity to the query. The key strategy was to devise the best weight algorithm to represent web pages and query in a vector space, so that closeness in such a space would be correlated with semantic relevance.

The basic procedure executed by any web crawling algorithm

takes a list of seed URLs as its input and repeat URLs as it input and repeatedly executes the following steps.

- Enter the query
- Download the corresponding web page
- Check the relevancy of the web page
- Extract any link contain in it
- Add these links back to the URL list
- After all URLs are processed, return the most relevance page.

Some of the web crawling algorithm used by crawler that will consider are:

The Fish-Search [4] is an example of early crawlers that maintain the order of web pages in queue according to the relevance factor. In the fish search, every keyword is assigned to 0 or 1.So disadvantage of fish search is, all keywords have same priority value. Fish Search is a dynamic heuristic search algorithm.

The Shark-Search [5] is an extended version of Fish-Search, in which assign to priority value to keywords is more than 0 and 1.

Info spiders and Best-First are additional examples of focused crawling methods. Info spider uses the neural networks and Best-First method uses VSM to compute the relevance between candidate pages and the search topic. Shark-Search crawlers may be considered as a type of Best-First crawlers, but the former has a more complicated function for computing the priority values. In, Best-First was shown most successful due to its simplicity and efficiency. N-Best-First is generalized from Best-First, in which N best pages are chosen instead of one. A relatively more recent type of focused crawlers adopts learning-based approaches to relevance prediction.

III. Proposed Work

In this section, we describe how to improve and retrieve more relevance web pages.

Basically ontology based web crawler actually keyword focused on web crawler. Ontology based web crawler calculated total count [6] of keywords which are store in database or keyword related to query during surfing of web pages and can retrieve the web page and calculate the relevance factor by multiply weight factor of keyword.

Relevance Factor =∑ total count of keyword * weight factor

This relevance factor [6] does not focus on the total number of words in web page. But total number of words should be affect the relevance factor .We can see that by example three web page w1,w2,w3 and minimum relevance factor is 4.

Table 1: Relevance Factor of Web Pages

Web Page	Total number of words(N)	Relevance factor1 (rf1)	Relevance factor2 =(rt1/N)
W1	300	6.00	0.002
W2	600	6.00	0.001
W3	1000	11.00	0.0011

According to ontology concept till now, frontier (queue) of relevant web page are as w3, w2, w1 that is not focus on total number of words. If we focus on the total number of words in web pages then frontier (queue) will be w1, w3, w2.

A. Advantage of Proposed Work

"Ontology based web page for data mining" calculated relevance factor by considering the total number of word that also affect the maintaining the frontier queue. This is following advantage.

- Retrieve more relevant web page.
- Crawling time will be reducing.
- Space will be saving.

B. Methodology

In our approach we crawl through the Web and add Web pages to the database, which are related to a specific domain (i.e. a specific ontology) and discard Web pages which are not related to the domain. In this section we will show how to determine domain specific page.

1. Relevance Calculation

In this section we describe our own algorithm depending on which we calculate relevancy of a Web page on a specific domain.

2. Weight Table

We want to add some weights to each term in the ontology. The strategy of assigning weights is that, the more specific term will have more weight on it. And the terms which are common to more than one domain have less weight. The sample Weight table for some terms of a given ontology of the table shown below:

Table 2: Weight Table for the Ontology

Ontology Terms	Weight
Mining	1.0
Cleaning	0.6
Integration	0.4
Selection	0.5

3. Relevance Calculation Algorithm

In this section we design an algorithm how relevance score of a Web page is calculated.

INPUT: A Web page (P), a weight table.

OUTPUT: The relevance score of the Web page (P).

Step 1: Initialize the relevance score of the Web Page (P) to 0. RELV_P=0.

Step 2: Select first term (T) and corresponding weight (W) from the weight table.

Step 3: Calculate how many times the term (T) occurs in the Web page P. Let the number of occurrence is calculated in COUNT.

Step 4: Multiply the number of occurrence calculated at step3 with the weight W.

Let call this Keyword_weight and Keyword_weight =COUNT * W.

Step 5: Add this term weight to RELV_P. So new RELV_P will be,

RELV_P = RELV_P + Keyword_weight

Step 6: Actual RELV_P=RELV_P/ count of total words.

Step 7: Select the next term and weight from weight table and go to step 3, until all the terms in the weight table are visited.

Step 8: End.

4. Relevance Calculation of a Web Page

In fig. 1 we have shown an example of the above algorithm. From the Figure we can see that the total relevance of the Web page is 4.4(3.0+1.0+0.4). Suppose total number of words is 100. Then actual relevance factor is 0.044. If our relevance limit is 0.04 then

we can add this page to our database as a domain specific page. This algorithm can be used as ranking algorithm also. If two pages P1 and P2 got relevance x, y respectively and x>y, then we can say that P1 is more important than P2.

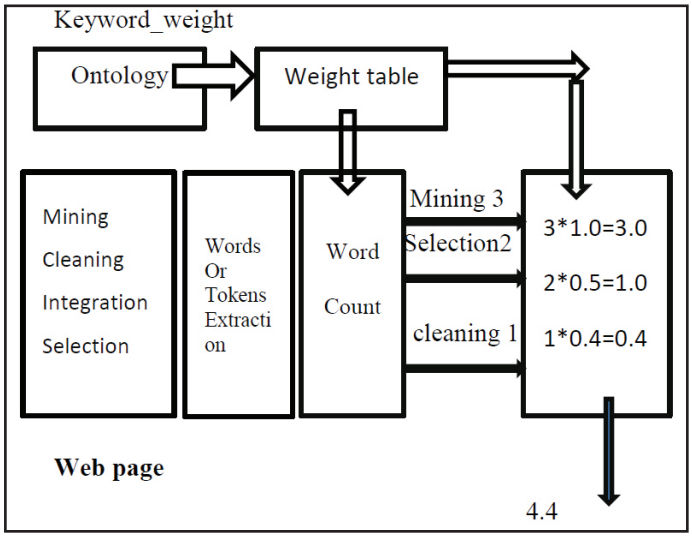

Fig. 1: Relevance Calculation of a Web Page

IV. Conclusion and Future Scope

A. Conclusion

Though Ontology based web crawler for data mining for searching the Data mining related topics are evolving constantly, the growth rate of these improvements is likely to be slight. Search engines based on a new concept as the domain specific search technology, are effectively able to handle the above mentioned problems.

Main advantage of using ontology web crawler over other web crawler is that it works intelligently, efficiently and doesn't need relevance feedback. It reduces number of extracted web pages thus it takes less time for crawling as it downloads relevant web pages only. Intension is to retrieve relevant web pages and discards the irrelevant web pages. We have developed an ontology based crawler using best knowledge path which fetches web pages according to relevancy decision mechanism. Below measurable advantages were found on comparing results with traditional crawlers.

1. Reduction in number of extracted searched web page URLs.
2. Reduction in turnaround time for crawling process.

This algorithm is potential and helpful to solve day to day troublesome for internet user and can save searching time.

B. Future Scope

Though we believe that our projected crawler takes care of everything an efficient crawler needs, there is still a window of improvement in our crawler that can be addressed. In our relevancy calculation algorithm, we have to set the weight of the ontology term manually. A mechanism can be devised such that after reading the ontology and after visiting certain Web pages it can provide the weight of the ontology term automatically. Also, the processing time of the Web crawler can be improved. In our algorithm the ontology remains static; ontology can be evolved dynamically by adding new concepts and relations while visiting Web pages.

References

[1] [Online] Available: https://en.m.wikipedia.org/wiki/Web_crawler

[2] Sergey Brin, Lawrence Page,"The Anatomy of a Large-Scale Hyper textual Web Search Engine", Computer Science Department, Stanford University, Stanford, USA, 1998.

[3] T.R.G ruber,"What is an Ontology?," [Online] Available: http://wwwksl.stanford.edu/kst/what-is-ontology.html

[4] De Bra P., Houben G., Kornatzky Y., Post R.,"Information Retrieval in Distributed Hypertexts, In: Proceedings of the 4th RIAO Conference, pp. 481 - 491, New York, (1994).

[5] Hersovici, M., Jacovi M., Maarek Y., Pelleg D., Shtalhaim M., Ur S.,"The shark-searchalgorithm - An application: Tailored Web site mapping", In Proc. 7th Intl. World-Wide Web Conference, Brisbane, Australia, (April 1998).

[6] [Online] Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4608260&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4608260

Swasti Singhal received her B.tech degree in I.T. from Uttar Pradesh Technical University, Lucknow and M.Tech in CSE from Amity University, Noida. Presently, she is working in Galgotia's College of engineering, Greater Noida and having 7 years of teaching experience. Her research interest includes Data mining and software testing.

Neeraj Kumar is pursuing his B.tech degree in Information Technology from APJ Abdul Kalam Technical University; Lucknow. His research interest includes Data mining and web technology.