A Prioritized Method of Searching a Keyword (Access over Internet) and Optimization

Amit Singh[1], Sonia Arora[2]

[1](PG Student): Dept. of Computer Science Engineering, World College of Technology and Management.
[2]Assistant Professor: Dept. of Computer Science Engineering, World College of Technology and Management.

Abstract: **The enormous content of information available on the World Wide Web makes it important topic for data mining research. Data mining techniques""s application to the World Wide Web is known as Web mining where this term has been used in three different ways; which are Web Content Mining, Web Structure Mining and Web Usage Mining. Web Mining uses the data mining techniques to automatically discover and extract information from web documents/services. It is used to discover useful information from the World-Wide Web and its usage patterns. Web Crawling, also known as Web spider, an automated indexer, an ant or a Web scutter is a program that browses the world wide web in a systematically, automated manner for indexing the content of web pages and keep the copy of all the pages that it has already visited for later processing. Web crawling is important to collect data for business intelligence, for market research about the services offered by the user and to determine and assess trends in a given market, to collect user behaviour information so that product can perform better and to develop a relevant product with more relevant contents. In my thesis work, I have explore data mining tool RAPIDMINER and showed how the operator of Crawling Web can be simulated into RAPIDMINER and result of Web Crawling can be generated accordingly. In my work, I have also designed an effective Web Search Engine in which we can give the lengthy query which will save the time of the user in searching the query. Web is expending day by day and people generally rely on search engine to explore the web. In such a scenario it is the duty of service provider to provide proper, relevant and quality information to the internet user against their query submitted to the search engine. It is a challenge for service provider to provide proper, relevant and quality information to the internet user by using the web page contents and hyperlink between the web pages.**

Keywords: **Optimization, RAPIDMINER.**

1. INTRODUCTION

Web Mining uses data mining techniques to automatically search and extract information from web services or documents. It is used to discover useful information from the World-Wide Web and its usage patterns. The subtasks of Web mining are:-

- Resource Finding: - To retrieve intended documents and services on Web.
- Information Selection/pre-processing:- Automatically extracting and pre-processing specific information from newly discovered Web resources.
- Generalization: - Refers to as the discovery of general patterns at individual Web sites and across multiple sites.
- Analysis: - The interpretation and/or validation of the mined patterns.
- Visualization: - Presenting the result of an interactive analysis in a visual, easy to understand fashion.

The Three Web mining categories depends on which kind of data to be mined that is mining for information or mining the Web link structure or mining the user navigation pattern. The process of Mining the information focuses on the development of various techniques to assist a user in searching documents meeting a certain criterion that is web content mining. It is the process of discovering useful information from the content of web pages, which includes image, text, audio, video etc. Web Link Structure Mining focuses on developing the different techniques to take benefits of the collective judgment of the quality of web page that is available in the form of hyperlinks known as web structure mining. Web structure mining discovers the model underlying the link structures of the web. Model is based on the topology of hyperlinks with or without description of links. Mining for user navigation patterns focuses on techniques which study the user behavior when navigating the web that is web usages mining. Web usage mining discovers the user access patterns from Web servers. And, Web usages data include data from the web server proxy server logs, browser logs, access logs, registrationdata, mouse clicks and scrolls, user session or transactions, user profiles, user queries, bookmark data, cookies, or any other data as result of interaction.

1. Web Content Mining

It discovers useful or potential information or knowledge from the contents of Web pages. Web content is very rich consisting of textual, images, audio, video etc and metadata (data about data) as well as hyperlinks. The data may be unstructured (free text) or structured (data from a database) or semi-structured (html) or Multimedia data (receive less attention than text or hypertext) although much of the Web is unstructured.

2. Web Structure Mining

It discovers the structure information from the Web. It generates the structural summary about the Web page and Web site and Discover the link structure of the hyperlinks at the inter-document level. It discovers the nature of the hierarchy of hyperlinks (i.e. inlink or outlink) in the website and its structure and Discovers similarities between sites. The subtasks of Web structure mining are:

- Finding Information about the Web page

It means to retrieve information to know about the quality and relevance of the Web page and find the authoritative on the topic and the content.

- Inference on Hyperlink

The Web page contains not only information but also hyperlink, which contains huge amount of annotation. Hyperlink identifies the endorsement of the author of the other Web page.

- Authority and Hub

A hub is a page that has link to many authorities and an authority is a page with good content on the query topic and pointed by many hub pages i.e. it is relevant and popular.

2. ALGORITHMS FOR WEB STRUCTURE MINING

1. HITS Algorithm

HITS that is Hyperlink-Induced Topic Search was proposed by Jon M. Kleinberg. The approach consists of two phases:-

- It collects a starting set of 200 pages using the query terms from the root set of pages which is known as index-based search engine. Including all the pages, the root set of pages is expanded into a base set to which the root set pages link , and all the pages that links to a page in the root set of pages.

- A phase i.e. Weight-propagation phase is initiated. It is determined by the numerical estimates of hub and authority weights using an iterative process. Associate a non-negative authority weight, ap, and a non-negative weight, hp, with each p in the base set and Initialize all a and h values to a uniform constant. The authority and hub weights are updated based on the following equations:

$ap = \sum(q \text{ such that } q \rightarrow p)\ hq$ (1)

$hp = \sum(q \text{ such that } q \leftarrow p)\ aq$ (2)

Equation (1) implies that if a page is pointing to by many good hubs, the weight of the authority should increase (means sum of current hub weights of all the pages that points to it). Equation (2) defines that if a page is pointing to many good authorities, the weight

of the hub should increase (means it is the sum of the current authority weights of all the pages it points to). Now, we define a matrix M. The rows and columns of Matrix M correspond to the Web pages having entry Mij=1 if page i of Web page links to page j of Web page, and Mij = 0 if not. Let a and h be the authority and hub weight vectors having ith component as the degrees of authority and hubbiness of the ith page. Then we have:

$h = M \times a.\ a = MT \times h.$

Where AT is matrix A‟s transposition.

2. Page Rank Algorithm

Page rank algorithm was presented and published by Sergey Brin and Larry Page at the Seventh International World Wide Web Conference (WWW7) in April 1998. It is a search ranking algorithm that uses hyperlinks on the Web. Based on this algorithm, the search engine Google was built, which has been a huge success. The following ideas based on rank prestige [86] are used to derive the Page Rank algorithm:

1. An implicit communication of authority to the target page is a hyperlink from a page that points to another page. So, the prestige of page i will be more if it receives the more in-links.

2. The in-link Pages of page i also have their own prestige scores. And A page that points to page i and have a higher prestige score is more important than a page with a lower prestige score that points to i. Or in other words we can say, a page is more important than the other pages if it is pointed to by many other important pages.

According to rank prestige in social networks, the importance of page i (i‟s Page Rank score) is determined by summing up the Page Rank scores of all pages pointing to i. Since a page may point to many other pages, the prestige score of this page should be shared among all the pages that it points to. Suppose Web as a directed graph G = (V, E), where V is the set of vertices or nodes that is the set of all pages and E is the set of directed edges in the graph that is hyperlinks. Let the total number of pages on the Web be n (that is, n =|V |). The Page Rank score of the page i (denoted by P(i))is defined by

$P(i) = \sum P(j)/Oj,$

(j,i)є E

Where Oj is number of out-links of page j.

the Web graph to satisfy the conditions, the following Page Rank equation is produced:

$P = (1 - d)e + dATP$

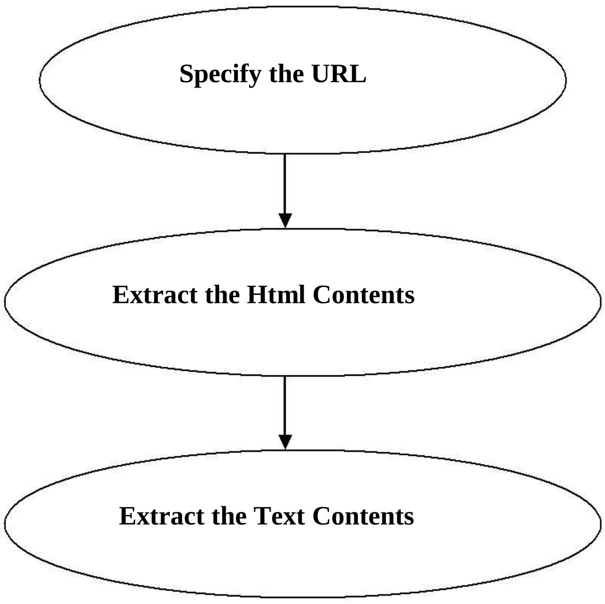
3. PROPOSED METHODOLOGY

Research Design

The complete research work will be performed in following steps.

Extract the Document

The first step of the research is to extract the Web Document. For the web document extract we will prefer some news site. We need to perform the web content mining to extract the document. The basic architecture followed by Web page extraction is given as



Query Summary Generation

To summarize the query and the web document to perform effective matching

1. Extraction of Keywords

2. The Frequency of the appearance

3. Stemming the keywords

4. Update frequency

4. ANALYSIS

In the final step of research the results will be analyzed. The analysis will be done in terms of:-

Proposed Approach

Summarization of query is a difficult task in text data mining owing to the high-dimensionality and sparse nature of text documents. It requires efficient algorithms which can address this high dimensional Summarization problem. Text Summarization plays an important role in web based applications and text data mining. Major applications of Document Summarization include .

Effective Search result

Search results we mean the documents returned in response to a query. Document Summarization is applied in web search engines (The automatic generation of a taxonomy of Web documents like that provided by Yahoo, Google etc.) to improve search results. Its benefit is more effective information presentation to the users.

Cluster-based effective navigation

This is an interesting alternative to “searching” keyword”, the standard information retrieval paradigm. This is extremely useful in cases where users prefer browsing over searching when they are not sure about which search terms to be use. Its benefit is provision of alternate user interface i.e. „search without typing“. The result of a query is now matched to a cluster rather than to each document thus reducing the search space.

Query Summarization Procedure

The standard query Summarization process consists of the following steps:

Pre-processing

The documents to be clustered are in an unstructured format therefore some pre-pre-processing steps need to be performed before the actual Summarization begins. The pre-processing includes Tokenization, Stemming of document words, and Stopword removal. Tokenization means tagging of words where each token refers to a word in the document.

Stemming involves conversion of various forms of a word to the base word. E.g. „computing‟ and „computed‟ will be stemmed to the base word „compute‟. Similarly „sarcastically‟ is stemmed to the word „sarcasm‟. The Porter‟s Algorithm is the most popular stemming technique for English Language documents. Snowball is a popular tool using this stemming algorithm.

Stop word removal: Stop words are the words present in documents which do not contribute in differentiating a collection of documents hence, are removed from the documents. These are basically articles, prepositions, and pronouns. Standard stoplists are available but they can be modified depending upon the kind of dataset to be clustered.

Feature Selection and Document Representation Model

Documents need to be represented in a suitable form for Summarization. The most common representation includes the recognition. which treats documents as a bag-of-words and uses words as a measure to find out similarity between documents. In this model, each document Di is located as a point in a m-dimensional vector space, Di = (wi1,wi2, . . .,wim), i = 1,. . .,n, where the dimension is the same as the number of terms in the document collection. Each component of such a vector reflects a term within the given document. The value of each component depends on the degree of relationship between its associated term and the respective ocument. There are three most common term weighting schemes to measure these relationships:

Application of Summarization Algorithm

The Summarization algorithm generates clusters based on similarity measure and data representation model. Numerous Summarization algorithms have been implemented in document Summarization literature which will be discussed ahead in the Previous Work section.

Flow Chart

There are differences in the working of various search engines, but all of these search engines perform three basic tasks, which are:

1. All of the search engine search the Internet or select pieces or part of the Internet based on important words given as input.
2. All of the search engine keep with them an index of the words they find on the internet, and where they find them.
3. All of search engine allow users to search for words or combinations of those words that are found in that index.

5. CONCLUSION

So, Web Crawling is important to utilize the time and to make the search faster. It saves the result, i.e., the pages crawled by web crawler in a folder indexed by the number so that it can be processed later if required. We can set the different parameters according to our requirements and we will get the result accordingly. In future, we will calculate the result by changing the various factors and will check whether they affect the performance of Web crawler, for e.g., overlapping of the web documents, quality of downloaded web documents, Network traffic problem and change of web documents.

In this present work, We have designed an effective Web Search Engine in which user can give the lengthy query or upload the query from the text file. Then the various analysis has done for testing. This will save the web users time to search the query.

REFERENCES

[1] http://en.wikipedia.org/wiki/Web_crawling

[2] http://www.encyclopedia.com/topic/Webcrawler.aspx

[3] http://en.wikipedia/wiki/Distributed_web_crawling

[4] http://www.wisegeek.org/what-is-a-web-crawler.htm

[5] www.mendeley.com/catalog/web-crawling

[6] Ioannis Katakis," Automated Tagging for the Retrieval of Software Resources in Grid and Web Infrastructures", 2012 12th IEEE/ACM International Symposium on Cluster, Web and Grid Computing 978-0-7695-4691-9/12© 2012 IEEE

[7] MARIOS D. DIKAIAKOS," Minersoft: Software Retrieval in Grid and Web Computing Infrastructures", ACM Transactions on Internet Technology, Vol. 12, No. 1, Article 2, Publication date: June 2012 ACM 1533-5399/2012/06

[8] Bernardo Ferreira," Management and Search of Private Data on Storage Webs", SDMCMM‟12, December 3-4, 2012, Montreal, Quebec, Canada. ACM 978-1-4503-1615-6/12/12

[9] Maria-Elena Hernandez," Synchronized Tag Webs for Exploring Semi-Structured Clinical Trial Data".

[10] AAMEEK SINGH," Search-as-a-Service: Outsourced Search over Outsourced Storage", ACM Transactions on the Web, Vol. 3, No. 4, Article 13, Publication date: September 2009. ACM 1559-1131/2009/09

[11] Venkateshprasanna H.M.," Enterprise Search through Automatic Synthesis of Tag Webs", COMPUTE „11, March 25-26, Bangalore, India ACM 978-1-4503-0750-5/11/03

[12] Yuko Arai," Query Log Perturbation Method for Privacy Preserving Query", ICUIMC"10, January 14-15, 2010, SKKU, Suwon, Korea. ACM 978-1-60558-893-3

[13] Hang Guo," Personalization as a Service: the Architecture and a Case Study",