A Review of Focussed crawling techniques in Web mining

Ravneet Kaur * and Er. Jyoti Arora**
M.Tech Pursuing*, Assistant Professor**
Desh Bhagat University, Mandi Govindgarh (Punjab)

Abstract
In this review, we have studied concepts and techniques of web mining technologies and its importance. We have reviewed two important web mining techniques and looked at their concept and general theory along with their pros and cons. Then we have compared the two based on few metrics and presented the findings.

INTRODUCTION
Web Crawlers And General Engines For Website Searching Are Facing A Big Problem In The Field Because Of The Advancements In The Web. The Computerized Tool Or The System Which Helps Us To Search And Browse Only The Relevant Things On The Internet Search Engines Is Focused Web Crawlers. Topical Crawling Only Pays Attention In Only The Topic Being Present But The Focused Crawling Also Works With Some Labeled Examples Of Relevant And Irrelevant Web Pages. A Focused Crawler Delivers Relevant Web Pages On A Given Topic In The Web. There Are A Number Of Problems Related To Existing Focused Crawlers, In Meticulous The Ability To “Tunnel” During Lowly Ranked Pages In The Search Path To Greatly Ranked Pages Related To A Topic Which Might Re-Occur Additional Down The Search Path. A Focused Crawler Has The Following Main Machinery:
(A) It Finds Out The Relevant And Irrelevant Pages From A Lot Of Pages.
(B) A Methodology Which Includes The Way About.


The Problem Is Tackled In Where Strengthening Knowledge Allows Praise Assignment Through The Search Process, And Therefore, Permitting Off-Topic Pages To Be Incorporated In The Search Path. Nevertheless, This Approach Requires A Large Number Of Training Example, And The Method Can Only Be Qualified Offline. In, A Bunch Of Classifiers Are Educated On Examples To Estimate The Distance Of The Current Page From The Closest On-Topic Page.
But the training formula is quite complex.

Our focused crawler aims at providing a simpler substitute for overcoming the issue that instantaneous pages which are lowly ranked associated to the topic at hand. The idea is to again and again execute an comprehensive search up to a given depth. Starting from the "relatives" of a greatly ranked page. Therefore, a set of candidate pages is obtain by retrieving pages reachable within a given border from a set of initial seeds. We get the best pages from the given set of pages which are included in the search. This page and its "relatives" get inserted into the set of pages from which to ensue the crawling process. Our supposition is that an "ancestor" with a good reference is liable to have other useful references in its offspring further down the ancestry even if abrupt scores of web pages closer to the forebear are low. We define a degree of relatedness with respect to the page with the greatest score. If \( r \) is large, we will include more far-flung "cousins" into the set of seeds which are supplementary and supplementary away from the highest scored page.

A web search engine is designed to search for information on the world wide web. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

Search engine optimization (seo) is the process of affecting the visibility of a website or a web page in a search engine's "natural" or un-paid ("organic") search results. Seo may target different kinds of search, including image search, local search, video search, academic search, news search and industry-specific vertical search engines.

As an internet marketing strategy, seo considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience.

A web crawler systematically browses the world wide web, typically for the purpose of web indexing. Web search engines and some other sites use web crawling to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. A focused crawler is a web crawler that collects web pages that meets precise property, by giving precedence to the crawl frontier and association the hyperlink examination process. Some anticipates may be based on simple, resulting and surface properties. Example is a crawler's aim is to crawl pages from only the known domain. An important page belongings pertains to topics, leading to topical crawlers. For example, a topical crawler deploys to collect pages about solar power, or swine flu, while reducing resources spent fetching pages on other topics. Crawl boundary administration may not be the only tool used by focused
Crawlers; They May Use A Web Fact List, An Web Text ,Index, Backlinks, Or Any Other Web Artifact.

A Focused Crawler Must Predict The Likelihood That An Unvisited Page Will Be Relevant Before Essentially Downloading The Page. A Possible Predictor Is The Anchor Text Of Links; This Was The Advance Taken By Pinkerton In A Crawler Developed In The Early Days Of The Web. Topical Crawling Was First Introduced By Filippo Menczer Chakrabarti Et Al. Coined The Term Focused Crawler And Used A Text Classifier To Prioritize The Crawl Boundary. Andrew Mccallum And Co-Authors Also Used Reinforcement Learning To Focus Crawlers. Diligenti 'Et Al. Traced The Context Graph Foremost Up To Relevant Pages, And Their Text Content, To Educate Classifiers. With The Vast Amount Of Information Available On The Web And About Millions Of Pages Available On The Web, It Becomes Challenging Task To Search All The Pages On The Web By Search Engines. Whenever An User Type A Word On Search Engines Like Google, Yahoo Etc., It Gives Millions Of Results To The User. But Most Of The Results Are Of No Use. Therefore, Focused Crawlers Were Introduced To Select Urls Which Are Relevant To A Topic. But Here Issue Is How To Predict Whether The Content Of The Urls Are Relevant Or Not. So, Here We Are Applying Data Mining Techniques To Improve Prediction & Performance.

The Aim Of A Focused Crawler Is To Seek Out A Subset Of The Web To Only Gather Documents On A Specified Topic Which Are Relevant And Quality Documents And Avoid Irrelevant Documents.


LITERATURE SURVEY

According To D. Saraswathi Et. Al.[1] The Web Is Both An Excellent Medium For Sharing Information As Well As Attractive Platform For Delivering Products And Services. This Platform Is, To Some Extent, Mediated By Search Engines In Order To Meet The Needs Of Users Seeking Information. Search Engine Has Enormous Amount Of Information And Return A Customary Answer For Given Query With Only Small Set Of Results. Most Of The Web Sites Are Interested To
Display The Web Pages Within Ten Search Results. In This Paper Author Has Designed A Spam Detector The Spam Detector Detects Whether The Links Are Spam Or Not Spam, If It Is Spam Remove From The Search Engine Results, Otherwise Display The Results To The User. Due To The Similarities Between Spam And Non-Spam The Existing Link Spam Identifiers Are Not An Effective Method To Classify The Web Links. Since Spam And Non-Spam Links Are So Similar, It Is Sometimes Very Difficult For A Human To Differentiate Between The Two. The Ideas And The Concepts Identified This Research Would Benefit The Users Who Search Information In The Search Engine. Since The Users Will Get Quality Web Links And No Need To Spend Much Amount Of Time To Search The Information In The Web.

According To Bireshwar Ganguley Et. Al.[2] Has Designed A Crawler, Which Is A System That Learns The Specialisation From Examples, And Then Evaluates The Web, Directed By A Relevance And Reputation Rating Mechanism. It Filters At The Data-Collection Level, Rather Than As A After-Processing Step. The Author Has Also Talked About Internet, Search Engines, Web Crawlers, Focused Crawlers And Block Partitioning Of Web Pages. In This Paper Approach Was To Partition The Web Pages Into Content Blocks. Using This Approach We Can Partition The Pages On The Basis Of Headings And Preserve The Relevant Content Blocks. The Information Can Be Used To Collect More On Related Data By Intelligently And Efficiently Choosing What Links To Go Behind And What Pages To Abandon. This Process Is Called Focused Crawling. A Focused Crawler Tries To Identify The Most Promising Links, And Ignores Off-Topic Documents.

According To M. Diligenti Et. Al.[3], According To Et. Al., Keeping Currency Of Search Engine Indices By Comprehensive Crawling Is Rapidly Becoming Unfeasible Due To The Growing Size And Lively Content Of The Web. Focused Crawlers Plan To Search Only The Division Of The Web Related To A Precise Group, And Tender A Possible Way Out To The Currency Difficulty. The Major Difficulty In Focused Crawling Is The Stage Suitable Acclaim Obligation To Different Credentials Along A Crawling Pathway, Such That Immediate Gains Are Not Accepted At The Expense Of Less-Obvious Crawl Pathways That Eventually Yield Superior Sets Of Important Pages. To Deal With This Dilemma We Present A Focused Crawling Algorithm That Makes A Scheme For The Context In Which Topically Relevant Pages Arise On The Web. This Context Scheme Can Incarcerate Typical Link Automates Within Which Important Pages Occur, As Well As Scheme Content On Credentials That Commonly Simultaneously Occur With Relevant Pages. This Algorithm Additionally Leverages The On Hand Capability Of Large Search Engines To Provide Fractional Reverse Crawling Abilities. This Algorithm Shows Noteworthy Presentation Advancements In Crawling Competence Above Average Focused Crawling.

According To Hong-Wei Hao Et. Al.[4], Topic Relevance Of Pages And Hyperlinks Is The Key Issue In Focused Crawling. In

FOCUSED WEB CRAWLING

Focused web crawling is a technique in web mining in which relevant information is mined using a general knowledge of interest and intelligent choice of links to follow or discard based on relevancy parameters. The task of crawling starts with a set of initial links known as seed pages (or seed links) that indicates the prior information about the content a user is interested in. The content of these seed pages are then evaluated for topic relevance based on keyword matching like basic features to advanced machine learning techniques. The hyperlinks found on the seed pages are further filtered. The filtered links are then followed to collect more information and so on. A page from which the link is extracted is called ‘parent page’ and the one to which the link points to is called ‘child page’ or a ‘target page’.

FISH SEARCH

A fish search algorithm only considers those seed pages that have content matching the given query and their neighborhoods. Query can be simple set of keywords or a regular expression. Neighborhoods are the pages pointed to by the seed page. A fish search algorithm follows the internet as a directed graph in which webpage is a node and hyperlink is an edge, so the search operation could be abstracted as a process of traversing directed graph.

In this method, for each node, we compute the relevancy of the page (or node) as relevant (1) or irrelevant (0). The algorithm maintains a list of URL of pages to be searched. Each URL has different priority and the URLs with high priority are located at the front of the list and are searched earlier. If a relative page is found, it stands for that the food has been found by the fish. The key problem of fish search algorithm is a very low differentiation of the priority of pages in the list.

SHARK SEARCH

An extension or modification of basic fish search is Shark Search which is different in many aspects to that of Fish Search. URLs of pages to be downloaded are put into a priority table by taking into account the linear combination of source page relevance, anchor text and neighborhood of the link on the source page and inherited relevance. The inherited relevance is calculated by multiplying page’s relevance score by a specified decay factor.

The improvement is that instead of the binary evaluation of relevance as seen in fish search (relevant (1) and irrelevant (0)), this algorithm generates a fuzzy score i.e. a score between 0 and 1. Second, it uses Vector Space model for calculating the relevance Score. The drawback is that this algorithm neglects information of link structure.

Conclusion

In this review, we have studied the general theory of Web Mining as a subset of Data mining technology. Further we have looked at the concept of Focused Web Crawling and seen two important algorithms in Focused Crawling aka. Fish Search Algorithm and Shark Search Algorithm.

We have also seen that Shark Search is an improvement over Fish Search Algorithm and we have seen that it takes into account
many factors to calculate a Fuzzy Relevancy Score instead of Binary Score as in case of Fish Search. Also, in Shark Search, the algorithm does not follows the Link Structure and thus pose a major drawback.

REFERENCES

