Text Mining based Web-Crawling and Search Optimization Techniques

Minal Chandurkar, Pooja Gaikwad , Snehdeep Thakare , Saurabh Kurhade

Bachelor of Engineering (B.E.)

Department of Computer Technology,

Rajiv Gandhi College of Engineering and Research, Wanadongri, Nagpur

E-mail:- snehdipt1@gmail.com

Abstract: **— In this paper we are trying to optimize the search results for the user according to the user behaviour. The proposed model is designed to achieve the high performance for determining relevant information to answer what users want. The model would be better than other pattern based models, concept-based models, and state-of the-art term-based models in the effectiveness. The proposed deploying method has better performance for the interpretation of discovered patterns in text documents. This deploying approach is not only promising for pattern-based approaches, but also significant for the concept based model.**

Keywords- Apriori Algorithm, Search Engine, Text Mining, Web Crawling.

INTRODUCTION

A search engine is computer software that is continually modified to avail of the latest technologies in order to provide an optimization search results. Each search engine does the same function of collecting, organizing, indexing and serving results in its own unique way, thus employing various algorithms and techniques, which are their trade secrets. In short, the functions of an optimization search engine can be categorized into the following broad areas: *First,* crawl the web and locate all web pages. *Second,* personalized the data. *Third,* Optimization results, so that when a user does a search. The Simple Search can be presented first. The paper also highlights the shortcomings of the most popular technique. Also, various alternatives to those flaws are discussed. This paper focuses on the current prevailing ranking techniques and tries to probe the conditions under which they can give more benefit. Also the various techniques are discussed in detail to present the flaws in this paper, and finds a way to achieve the ideal technique. In this paper the optimization of search results is done on the basis of personal behaviour of the user.

We are trying to optimize the search results for the user according to the user behaviour. The proposed model is designed to achieve the high performance for determining relevant information to answer what users want. The model may be better than other pattern based models, concept-based models, and state-of the-art term-based models in the effectiveness.

The proposed deploying method has better performance for the interpretation of discovered patterns in text documents. This deploying approachis not only promising for pattern-based approaches, but also significant for the concept based model.A Web Search Engine is a software that is used to search information on the World Wide Web [1-4]. The information may be a specialist in web pages, images, information and other types of file.In this paper we are trying to optimize the search results for the user according to the userbehaviour. The proposed model is designed to achieve the high performance for determining relevant information to answer what users want. The model would be better than other pattern based models, concept-based models, and state-of the-art term-based models in the effectiveness.

The proposed deploying method has better performance for the interpretation of discovered patterns in text documents. This deploying approach is not only promising for pattern-based approaches, but also significant for the concept based model.

Most of the earlier work on optimize search was dedicated to increasing efficiency by finding out more relevant pages in less time i.e. to give more page-hit in less time. We needed a mechanism to check as to *What degree is a page relevant to the user'.* Thus it is needed that we should generate more result and more relevant information in given time. If the Link Based

checking is performed, it is seen that sometimes results are found but there may be times that the results are not found on the pages on which the links match to the users query and thus, such pages have to be ignored. Along with link checking content of the page has to be checked so as to determine the degree of content relevancy in a particular page. For this, the tags of the page are to be assigned weights (the mechanism that we call *parsing).* The results of the link based search and content based search has to be combined so as to find the total weight of the page and this will thus decide the relevancy degree of that page. The pages which will not have any weight will thus be useless for users and thus are deemed Irrelevant.

- ***Apriori Algorithm-is*** an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

- ***Text Mining-Text*** mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want.

- ***Web Crawling-Focused*** crawler [1] is used to selectively collect smaller Web pages collections according to a particular topic with high precision. A focused crawler will try to predict whether a target URL is pointing to a relevant Web page before actually fetching it. Focused crawlers rely on two kinds of algorithm to keep the crawling process on the track. First, Web analysis algorithm will evaluate the quality and relevance of Web pages pointed by target URLs. Second, Web searching algorithm will determine the optimal order in which the targets URLs are visited.

LITERATURE SURVEY

Ning Zhong and et al give many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. [1]

Ning Zhong and et al give the hypothesis that pattern-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance. [1]

Banu Wirawan Yohanes and et al give the size of the Web continues to grow, searching it for useful information has become more difficult. Focused crawler intends to explore the Web conform to a specific topic. This paper discusses the problems caused by local searching algorithms. [2]

Banu Wirawan Yohanes and et al proposed the genetic algorithm is used to optimize Web crawling and to select more suitable Web pages to be fetched by the crawler. Several evaluation experiments are conducted to examine the effectiveness of the approach. The crawler delivers collections consist of 3396 Web pages from 5390 links which had been visited, or filtering rate of Roulette-Wheel selection at 63% and precision level at 93% in 5 different categories.[2]

Bin Jiang and et al give one of the essential tasks in mining uncertain data, posts significant challenges on both modelling similarity between uncertain objects and developing efficient computational methods. The previous methods extend traditional partitioning clustering methods like k-means and density-based clustering methods like DBSCAN to uncertain data, thus rely on geometric distances between objects. Such methods cannot handle uncertain objects that are geometrically indistinguishable.[3]

Users' search behaviour has a strong effect on SEO, since many users click the first listing in a SERP. On the other hand, gaining a #1 SERP position for a particular search term does not guarantee that a user will click that link; they might click the #2 link or the #10 link if those page titles, descriptions, or URLs are more compelling. In the same vein, there is no clear evidence that users who click the first record will take further action; many SEO practitioners believe the more motivated users will scan an entire SERP before deciding which link to click. Regardless, user behaviour while visiting SERPs affects SEO decisions as well as the search engines' goal of relevance. [4]

PROPOSED MODEL

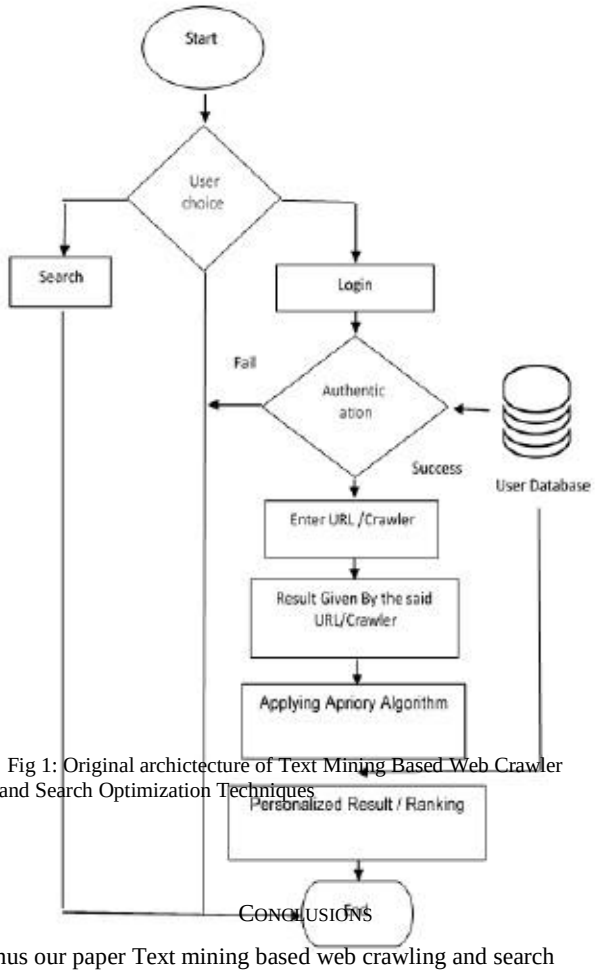

Fig 1: Original archictecture of Text Mining Based Web Crawler and Search Optimization Techniques

CONCLUSIONS

Thus our paper Text mining based web crawling and search optimisation technique is complete its Module 1 which is GUI i,e. user registration and login form and database is made for the this user registration and admin. In this paper, the search is optimized by using personalized way. The process is done by getting user Behaviour and give optimized search which saves the time of users and give the personalized output as required. The proposed work of this paper is fulfilled.

ACKNOWLEDGMENT

We would like to thank our guide Mr. Piyush Dhule for providing us with valuable resources and insights needed for my paper. He has been a guideline for us throughout the paper. We really appreciate his valuable time spent for the betterment of this paper. We would like to thank all our professors from the Computer Technology Department who have helped enrich our knowledge and for all their support and encouragement.

REFERENCES

[1]Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[2] Bin Jiang, Jian Pei, Yufei Tao,and Xuemin Lin,"Clustering Uncertain Data Based on Probability Distribution Similarity", TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013

[3] A. N. Langville and Carl Dean Meyer. Google's PageRank and Beyond : The Science of Search Engine Rankings, pp. l-19. Princeton University Press, 2006.

[4] Page, *et al.* The PageRank citation ranking: Bringing order to the Web, 1999

AUTHORS PROFILE

Mr. Piyush Dhule is a lecturer of Computer Technology Department at Rajiv Gandhi College of Engineering & Research, Nagpur, Maharashtra, India.

Ms. Minal Chandurkar is currently in her Final Year and pursuing her Bachelor in Engineering (B.E) Degree from Rajiv Gandhi College of Engineering and Research, Nagpur.

Ms. Pooja Gaikwad is currently in Final Year and pursuing hir Bachelor in Engineering (B.E) Degree from Rajiv Gandhi College of Engineering and Research, Nagpur.

Mr. Snehdeep Thakare is currently in Final Year and pursuing his Bachelor in Engineering (B.E) Degree from Rajiv Gandhi College of Engineering and Research, Nagpur.

Mr.Saurabh Kurhade is currently in his Final Year and pursuing his Bachelor in Engineering (B.E) Degree from Rajiv Gandhi College of Engineering and Research, Nagpur.