OPINION IDENTIFICATION FROM WEB DOCUMENTS

Amruta Sankhe[*]

Divya Racha*

Vijaya Sagvekar*

Abstract:

Generally, search engine retrieves the information using PageRank, Distance vector algorithm, crawling, etc. on the basis of the user's query. But it may happen that the links retrieved by search engine are may or may not be exactly related to the user's query and user has to check all the links to know whether the needed information is present in the document or not, it becomes a tedious and time consuming job for the user. Our focus is to cluster different documents based on subjective similarities and dissimilarities. Our proposed tool 'Web Search Miner' which is based on the concept of user opinions mining, which uses k-means search algorithm and distance measure based on Term frequency & web document frequency for mining the search results. It takes an opinion from the user on the results given by search engine in different web documents. in response to the user's multiple queries by downloading the pages in background, which saves the user's time of searching a particular information and it gives the best results for the precise search by giving a mined search links.

Keywords: web mining, opinion mining, and k-means clustering, web search.

[*] Atharva College of engineering, Mumbai University,India.

INTRODUCTION

Opinion mining refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials. Opinions are so important that whenever one needs to make a decision, one wants to hear other's opinions. This is true for both individuals and organizations. The technology of opinion mining thus has a tremendous scope for practical applications [1].

Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested multiple queries [2].

Text mining is directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire Web to retrieve the cluster content triggering the scanning of specific Web pages within those clusters. Though, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

Web text mining is very effective when used in relation to a content database dealing with specific topics. For example online universities use a library system to recall articles related to their general areas of study. This specific content database enables to pull only the information within those subjects, providing the most specific results of search queries in search engines. The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information [2].

This tool is imperative to scanning the many HTML documents, images, and text provided on Web pages. The resulting information is provided to the search engines in order of relevance giving more productive results of each search query.

The main aim of the Application is to identify and extracts subjective information in source materials. Search engine retrieves the various links relevant to the query entered by user. Sometimes it may happen that the user is not getting the satisfying results from the search engine. The main problem is that the search engine retrieves the data using Distance vector algorithm and Page Rank algorithm. To get the satisfactory result it is required that user knows the searching techniques, which every user doesn't know. Here the Opinion Miner comes in the picture; it gives the opinion on the retrieved result which guides the user about whether to go forward or not to go forward. The opinion given is by the user which provides filters to the system for performing text mining.

LITERATURE SURVEY

The rapid growth of World Wide Web in recent years has made it important to carry out resource discovery. Topic-specific web crawler collects relevant web pages of interested topics from the Internet, there are many relevant researches focusing on topic-specific crawling. However few works detail the topic-specific crawling with the user interests [3].

A lot of work has been done on clustering of texts which has already found practical applications like the clustering of search results. The topic of the clusters remains usually implicit in these approaches, though it would of course be possible to apply any keyword extraction algorithm to the resulting clusters in order to find characteristic terms. Like the work presented in this paper, Li and Yamanishi try to find characterizations of topics directly by clustering keywords using a statistical similarity measure. While very similar in spirit, their similarity measure is slightly different from the Jensen-Shannon based similarity measure used by, Moreover, they try to find the predominant overall topic of a whole text, while we are doing document level clustering based on subjective similarities and dissimilarities to give an opinion on document[4].

[5] This paper describes the research about Web data mining using Natural Language Processing. System accepts arbitrary data as input from Web document and then extracts information from the document. A new method to implement Web data mining is proposed in this paper. There are three steps in this system. First, the Web document will be decomposed to paragraph, sentence and phrase level. Second, extract information from all sentences. Finally, add the information to the knowledge model. The methods used have proved to be efficient for Web data mining with the experimental corpus.

As the addictive use of computers and 3G high speed internet have taken place in our day to day life. There is lots of information available on internet, some of them are structured, and some of them are unstructured.

There are three types of opinion mining, first one is Document Level opinion mining in which, the whole document is written about only one product and only by one person Next is Feature Level opinion mining, in which all the features or attributes are separated and for particular feature the opinions are extracted. It is too complicated so that is also not the focus of this paper. And the last is Sentence Level opinion mining, in which different people who have already used product, have written their opinions for product. This is the focus of this paper as it is interested in knowing different peoples' opinions. There are three techniques to used Naïve Bayesian algorithm because as this paper is focusing on supervise approach [6].

SYSTEM DESIGN

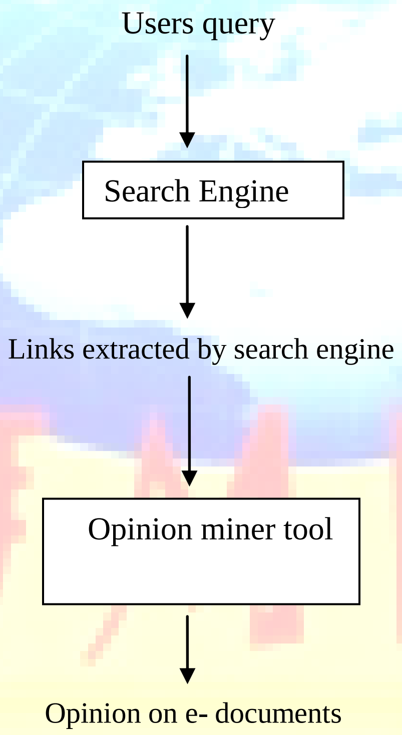


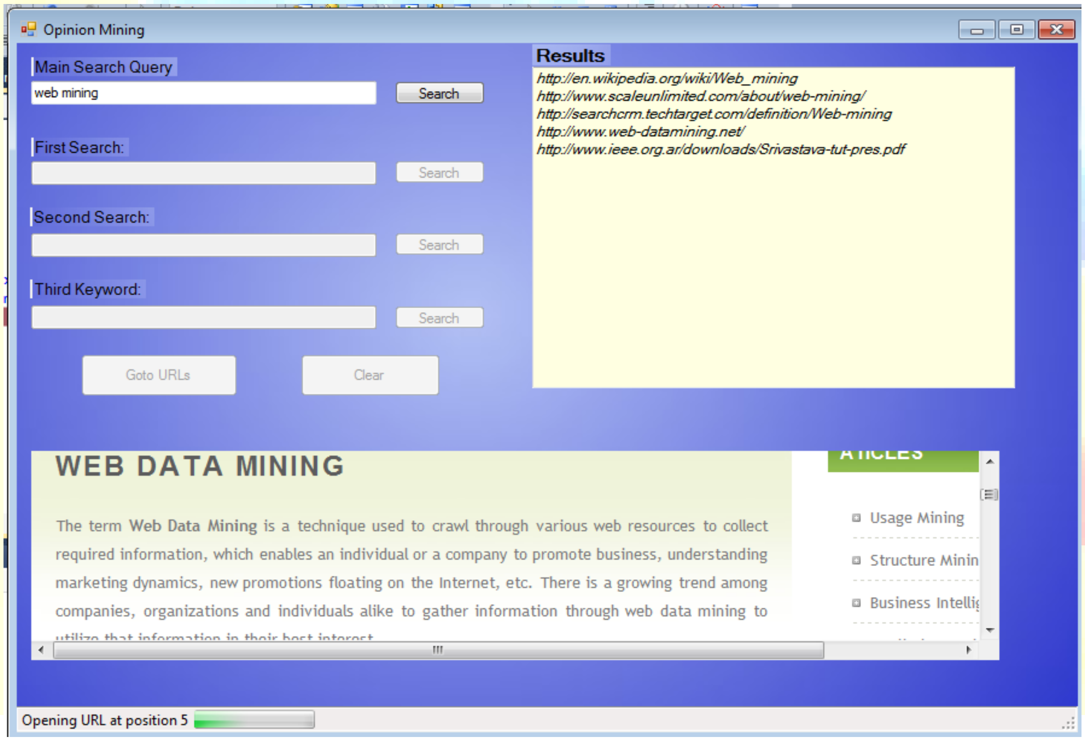
Fig .1 Design Model

Opinion miner can act as an additional feature to the search engine. The input to our tool is multiple e-docs retrieved by search engine on user's query. Our opinion miner is providing user the opinion on what these multiple e-doc says? How these docs are useful for the searcher as an opinion. Flow of the system will go as follows:

1. Input to system is multiple e-docs extracted by a search engine on user query on particular topic
2. Tool will search for the keywords entered by user and calculates its term frequency in each document.
3. Arrange these clusters for every document.
4. Multiple clusters from all documents are aggregated together.
5. Final opinion on all e-documents is presented in the form of most relevant links.

RESULTS AND DISCUSSIONS

The results of proposed system are discussed in this section.



Main Search Query: web mining

First Search:

Second Search:

Third Keyword:

Results

http://en.wikipedia.org/wiki/Web_mining
http://www.scaleunlimited.com/about/web-mining/
http://searchcrm.techtarget.com/definition/Web-mining
http://www.web-datamining.net/
http://www.ieee.org.ar/downloads/Srivastava-tut-pres.pdf

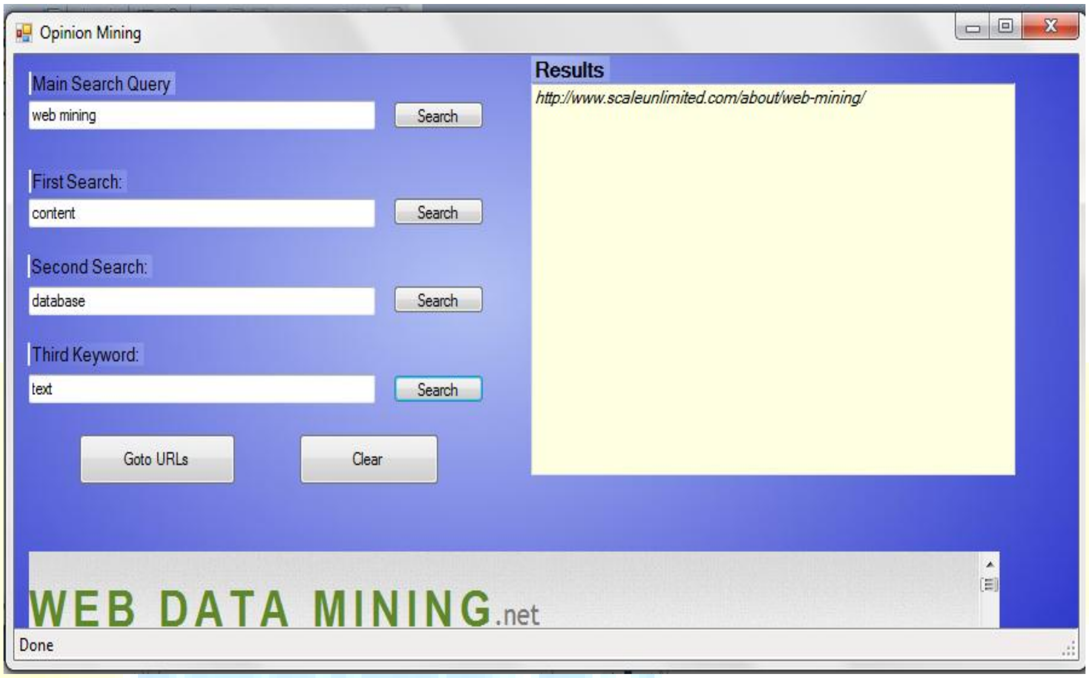
WEB DATA MINING

The term **Web Data Mining** is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest

Opening URL at position 5

Fig 2 putting first keyword

In the above figure, primary search query is executed. The resulting URLs are displayed as seen. User can view the website contents as the text is being mined by clicking on the desired URL in the result box.



Main Search Query: web mining

First Search: content

Second Search: database

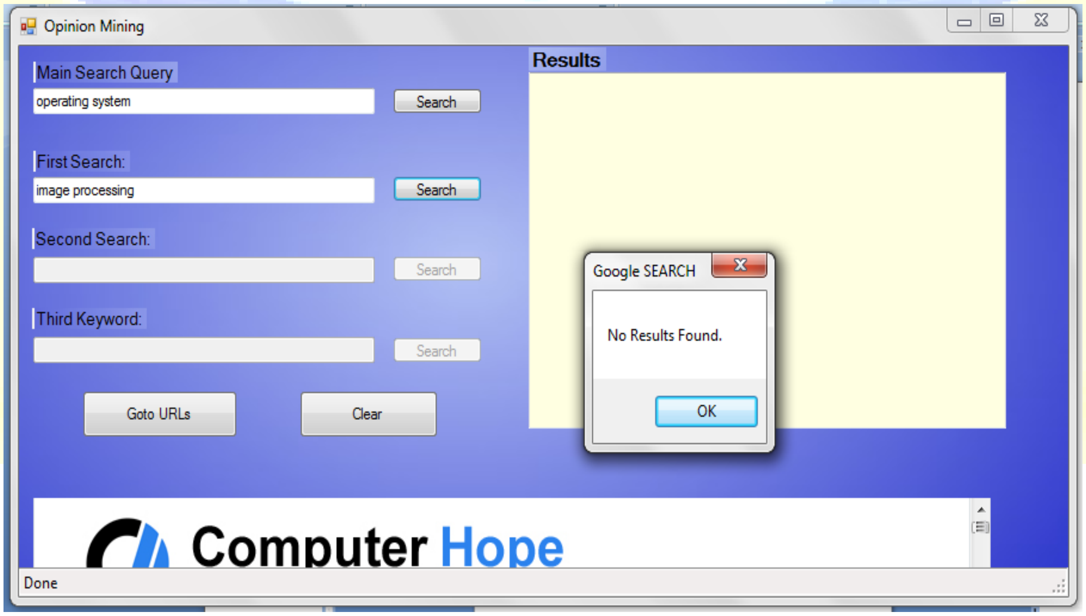
Third Keyword: text

Results

http://www.scaleunlimited.com/about/web-mining/

Fig3 filtration results

The final URLs are displayed after filtration of results. This is the final result that gives the URLs that have the specified keywords.



Main Search Query: operating system

First Search: image processing

Second Search:

Third Keyword:

Results

Google SEARCH

No Results Found.

Fig4 result for unrelated documents

The above figure shows the “no result found” dialog box when the application cannot find any documents.

CONCLUSION

We are giving opinion only on .pdf and .doc files by using this tool. Further we can integrate it with search engine so, it will not be bound to only one person and facility will be used by any user. Quick search will make searching job easier. Till the day the development is done is giving an opinion on blogs, comments, politics, etc. But nobody has drawn their attention to the e-documents, which is very important in the today's world, whatever information available on the internet is in the form of e-documents only. So this application draws its attention to e-documents and gives opinion on them which proves itself a very useful for the user for getting the fast result.

FUTURE SCOPE

Here we are filtering the pages on user opinion for the web pages. For which we have used Google search engine. In future we can integrate this on to various other search engines like Yahoo, etc. Quick search will make searching job easier. Further we can connect all instances of the application using a centralized sever. So that users can view what other users search and ultimately helps to make a decision.

REFERENCES

[1] Dr. A. Padmapriya S. Maheswaran," "Opinion Search and Retrieval from WWW,"IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 1, Issue 2 (May-June 2012), PP 13-17

[2] Sheetal Chouhan,Manish Shrivastava and Kavita Deshmukh,"A Noble Approach of Web Log Mining" VSRD-IJCSIT, Vol. 2 (7), 2012, 590-596

[3] Lei Xiang and Xin Meng," A Data Mining Approach to Topic-Specific Web Resource Discovery," Intelligent Computation Technology and Automation, 2009. ICICTA '09.

[4] Hang Li and Kenji Yamanishi," Topic Analysis Using a Finite Mixture Model," Association for Computational Linguistics,"2003

[5] Lei Xiang and Xin Meng, "A Data Mining Approach to Topic-Specific Web Resource Discovery , IEEE in, conf on Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference, October 2009

[6]Bing Liu "Web Data Mining", Springer Berlin Heidelberg 2011, pg 459-526

[7] Yue Chen, "Natural Language Processing in Web data mining", IEEE in Web Society (SWS), 2010 IEEE 2nd Symposium, Beijing, Aug. 2010