Search Utility for Domain Specific Search

[1]Amit Saroj, [2]Vaishali Raskar
[1,2]K J Somaiya Institute of Engineering & IT, Sion, Mumbai, Maharashtra, India.
[1]amit.saroj@somaiya.edu, [2]vaishali.v@somaiya.edu

Abstract : As we all know that vast amount of data is distributed over the World Wide Web so to search domain related data become tedious job. It becomes increasingly difficult to find just what we want. When we know that we want information of certain type, or on certain topic 'Internet search utility for domain specific search' application can be a powerful tool for searching the information related to particular domain for e.g. Search application for finding the relevant research papers. The purpose of this project is to develop an application for Information Poor People who have the lack of knowledge of inserting the appropriate query and finding the relevant document. This project will help the student, researchers, teachers and many other people to read the information that are available and to help the people to search relevant data on internet.

Keywords — Mosaicing, ransac, sift, featured pixel.

I. INTRODUCTION

Domain Specific Search application helps the user to search the relevant data by using machine learning technique to classify the data content and the information. This application aims at indexing the relevant pages from the internet and classifying the data appropriately by using the filters. When the user insert the query in the application with the help of query generator user get the suggestion if the user if unable to type the query. After that the query is send to the specific search engine proper indexing help to retrieve the link.

By using naive Bayesian text classifier algorithm extracts the content fields. For e.g. research paper. Which is used to search the Research papers i.e. IEEE paper and journals. With the Naive Bayesian algorithm we can extract the fields such as title authors and affiliation from the research paper headers and also the bibliography. These documents are converted to plain text and further processed if they are determined to be research papers (e.g. having Abstract and Reference sections). If the document found to be relevant the pages are returned to the user .The pages that are been returned to the user are sorted content wise, year wise. The appropriate graphical user interface help the user to get the data easily without going the certain search engine and finding the information which is time consuming and tedious job. Thus this application helps the people, student, teachers to the search the relevant data and without wasting time. The amount of data provided within short span of time. Also this application helps the people who don't have much searching knowledge.

II. LITERATURE SURVEYED

a. Focused Crawler Domain Specific Search

A focused crawler is domain specific web search may be described as a crawler which returns relevant web pages on a given topic in traversing the web. There are a number of issues related to existing focused crawlers, in particular the ability to ``tunnel" through lowly ranked pages in the search path to highly ranked pages related to a topic which might re-occur further down the search path. Focused crawler, which is described by two parameters, viz., degree of relatedness, and depth. Both provide an opportunity for the crawler to ``tunnel" through lowly ranked pages. There are two types of web crawling strategies deployed by focused crawler viz., breadth first search strategy

and ``best" first search strategy. Breadth first search strategy endeavors to build a general index of the web covering any conceivable topic by endeavoring to search a significant portion of the web.

b. Memex Domain Specific Search

Memex domain specific search model has been successful commercially. It does not work well for many government use cases but it still remains largely manual process that does not save sessions, requires nearly exact input with one-at-a-time entry, and doesn't organize of aggregate results beyond a list of links.

III. PROPOSED SYSTEM

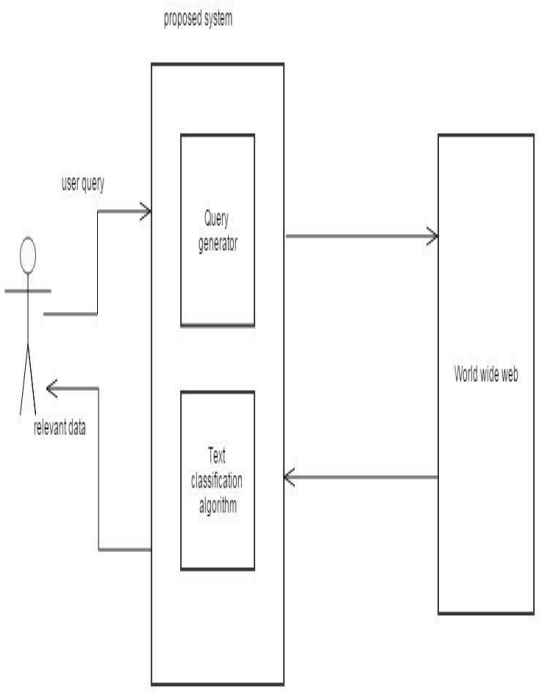

Fig. 1 Proposed system for Domain Specific Search

Internet Search Utility for domain specific search is an application which is used to search the relevant document .As the proposed block diagram state that when the user enter the query with the help of the query generator the user is provided with the suggestion which helps him to enter the appropriate query .After that query is send to the internet world wide web i.e. the internet over there the indexing is been performed. Indexing thus retrieve the appropriate links as per the entered query thus the indexing is achieved.

Once the link is been retrieved, by using Bayesian text classification algorithm the content is been filtered for e.g. Research paper. The algorithm find the internal content of the text i.e. the title author reference, abstract and the bibliography within spam of short period of time. if the document doesn't contain the internal entity the document will be discarded and the document with the proper internal entity will be processed in a plain text and the document will be returned to the user .Thus the system aim in providing the relevant document over the vast data thus it is not time consuming and very easily the document can be retrieved without the tedious job.

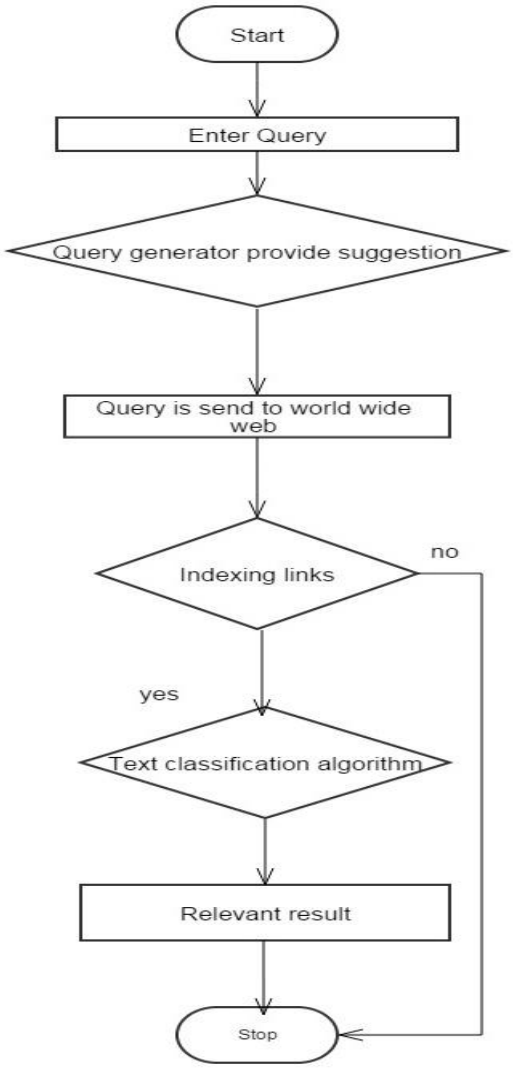

Fig. 2 Algorithm for Domain Specific Search

IV. BAYESIAN TEXT CLASSIFICATION

Naive Bayesian classification to the document classification problem. Consider the problem of classifying documents by their content, for example into spam and non-spam e-mails. Imagine that documents are drawn from a number of classes of documents which can be modeled as sets of words where the (independent) probability that the i-th word of a given document occurs in a document from class *C* can be written as

$$p(w_i|C)$$

(For this treatment, we simplify things further by assuming that words are randomly distributed in the document - that is, words are not dependent on the length of the document, position within the document with relation to other words, or other document-context.)

Then the probability that a given document *D* contains all of the words w_i, given a class *C*, is

$$p(D|C) = \prod_i p(w_i|C)$$

The question that we desire to answer is: "what is the probability that a given document D belongs to a given class *C*?" In other words, what is $p(C|D)$?

Now by definition

$$p(D|C) = \frac{p(D \cap C)}{p(C)}$$

and

$$p(C|D) = \frac{p(D \cap C)}{p(D)}$$

Bayes' theorem manipulates these into a statement of probability in terms of likelihood.

$$p(C|D) = \frac{p(C)}{p(D)} p(D|C)$$

Assume for the moment that there are only two mutually exclusive classes, *S* and ¬*S* (e.g. spam and not spam), such that every element (email) is in either one or the other;

$$p(D|S) = \prod_i p(w_i|S)$$

and

$$p(D|\neg S) = \prod_i p(w_i|\neg S)$$

Using the Bayesian result above, we can write:

$$p(S|D) = \frac{p(S)}{p(D)} \prod_i p(w_i|S)$$

$$p(\neg S|D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i|\neg S)$$

Dividing one by the other gives:

$$\frac{p(S|D)}{p(\neg S|D)} = \frac{p(S) \prod_i p(w_i|S)}{p(\neg S) \prod_i p(w_i|\neg S)}$$

Which can be re-factored as?

$$\frac{p(S|D)}{p(\neg S|D)} = \frac{p(S)}{p(\neg S)} \prod_i \frac{p(w_i|S)}{p(w_i|\neg S)}$$

Thus, the probability ratio p(*S* | *D*) / p(¬*S* | *D*) can be expressed in terms of a series of likelihood ratios. The actual probability p(*S* | *D*) can be easily computed from log (p(*S* | *D*) / p(¬*S* | *D*)) based on the observation that p(*S* | *D*) + p(¬*S* | *D*) = 1.

Taking the logarithm of all these ratios, we have:

$$\ln \frac{p(S|D)}{p(\neg S|D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i|S)}{p(w_i|\neg S)}$$

This technique of "log-likelihood ratios" is a common technique in statistics.

V. CONCLUSION

To search research papers through the internet is tedious job through this application the problem can be solved of the searching the relevant research paper at one place without traversing to various pages. Hence it saves times and helps the people those who don't have much searching or traversing knowledge of World Wide Web. Thus it helps the student the research the relevant papers sufficiently.

VI. REFERENCES

[1] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for Transfer Learning," Proc. 24th Int'l

Conf. Machine Learning (ICML ’07), pp. 193-200, 2007.

[2] Z. Cao and T. Yan Liu, “Learning to Rank: From PairwiseApproach to Listwise Approach,” Proc. 24th Int’l Conf. Machine Learning (ICML ’07), pp. 129-136, 2007

[3] H. Shimodaira, “Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function,” J. Statistical Planning and Inference, vol. 90, no. 18, pp. 227-244, 2000.

[4] J. Yang, R. Yan, and A.G. Hauptmann, “Cross-Domain Video Concept Detection Using Adaptive Svms,” Proc. 15th Int’l Conf. Multimedia, pp. 188-197, 2007

[5] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore, (2004) ”A Machine Learning Approach to Building Domain-Specific Search Engines

[6] Blitzer. J, Mcdonald. D, Pereira. R, (July 2006) “Domain Adaptation with Structural Correspondence Learning”, Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP ’06), pp. 120-128.

[7] Bo Geng, Linjun Yang, Chao Xu, and XianSheng Hua, (April 2012) “Ranking Model Adaptation for Domain-Specific Search”, IEEE Transaction on Knowledge and Data Engineering, vol 24, No. 4.

[8] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, (Apr 26,2012) ”Ranking Model Adaptation for Domain-Specific Search”, United States Patent Publications

[9] Burges. C.J.C et al Shaked. T, Renshaw. E, Lazier, A,Deeds. M, Hamilton. B, and Hullender. G, (2005), “Learning to Rank Using Gradient Descent”, Proc. 22th Int’l Conf. Machine Learning (ICML ’05).

