Middle-East Journal of Scientific Research 24 (S1): 220-223, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.S1.43

Winsome a Search Engine-Advanced Search

D. Keerthika and G. Sangeetha

Department of CSE, Valliammai Engineering College, Kanchipuram, India

Abstract: Now-a-days finding a direction during travel is made easy with the use of maps. All smart phones are associated with map applications like Google Maps, Bing Maps, MapQuest, Mappy and Here. If the search results Categorized in a manner then user easily find the results rather than spending time in other links. This type of feature in a search engine is not furnished by all search engine providers. To overcome this issue, a work is proposed to provide categorized results for the user to make the search easier and map feature is also added in this work to explore the places.

Key words: Search engine · Web crawler · Indexing · Searching · Mapping

INTRODUCTION

A Search engine is a software application which is devised to inquiry data or else information over World Wide Web. The input is given in the form of queries and output is offered as a response in the form of web pages. The responses may be in the form of text, audio, video, image, maps etc…A few search engines fetch data from database or open directories and preserve information by executing an algorithm using web crawler [1]. The process done by search engines is given as follows:

- Web Crawling
- Indexing
- Searching

Web Crawling: Crawling is done by the web crawlers. These are programs that maneuver the graph anatomy of the web to change from page to page. Other names for web crawlers are spiders, bots, robots, wanderers, worm etc…The primary function of web crawler is to redeem web pages and add them to the repository [2]. When all pages are redeemed there will be never needed crawling. The attributes of crawler are user-profiles, properties of fetchedpages and queries. There are two types of

Crawlers:

- Sequential Crawlers
- Multithreaded Crawlers

The Sequential crawlers works by following the proper order whereas the multithreaded crawler every thread follows a crawling loop. The advantage of multithreaded crawler is fast-speed and efficient use of bandwidth. Both crawlers are implemented using empty frontier. The architecture of web crawler [3] is shown in Fig.1.

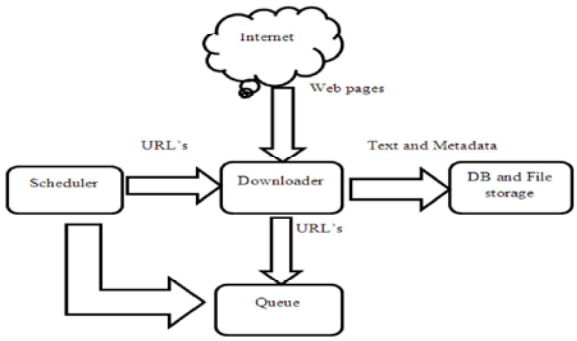

Fig. 1: Architecture of web crawler

Indexing: Indexing is the process of gathering, parsing and storing data to expedite speed and actual information retrieval (IR).Indexing is carried out to improve the performance of the search query [4]. without proper indexing and Search engine will have time complexities. For ex. With indexing we can scan 100 documents in a second but without that it will take 14 hours. It consists of size, lookup speed, storage, fault tolerance. It comprises clone of each crawler page. The indexing process is clearly depicted in fig.2.

Corresponding Author: D. Keerthika, Department of CSE, Valliammai Engineering College, Kanchipuram, India.

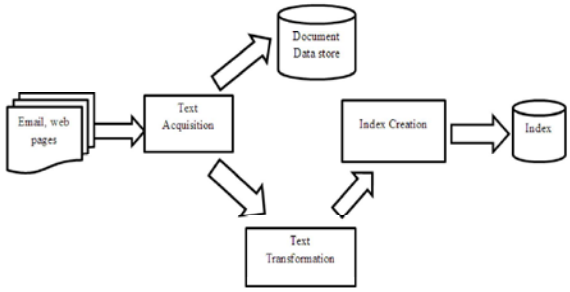

Fig. 2: Indexing process

Searching: Every Search Engine has its own searching methodology and ranking algorithms. The objective of searching is to produce best results efficiently. The search engine first gets the query and parses the input. It involves preprocessing, stemming and indexing. Preprocessing is cleaning the data. Stemming is removing the key suffixes. Then the data will be searched in the repository by the crawler and results will be displayed in the screen to the user as a search result.

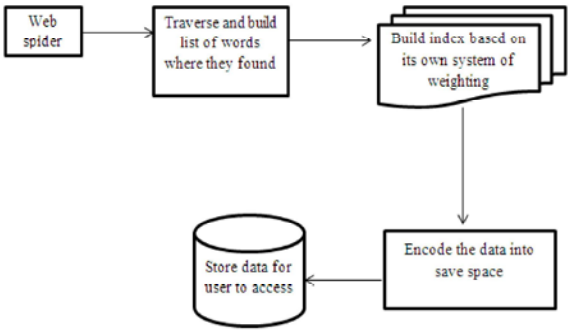

Fig. 3: Search Engine working

Contribution: The main objective of this work is to provide categorized results and map search. It will make the user's search easier and efficient to find the information. In the existing system, categorized results are acquired using different techniques. In this paper, we use Naïve Bayes theorem to provide the categorized results. And the next work is map search. In order to find the geographic location map search is implemented in this search engine.

Organization: The Remaining paper will proceed as follows. In section 2, we briefly explain about the System architecture. In section 3, we briefly explain about the implementation. This includes the algorithms with their steps. In section 4, we provide the results and discussion.In section 5, we provide the conclusion and discuss about the future works.

System Architecture: The search engine is one of the popular things in everyday life. The system architecture of this search engine is shown in fig.2.It consists of two main modules. They are:

- Categorized results
- Map search

Categorization is the method in which thoughts and entities are recognized, distinguished and tacit. It implies that entities are grouped into classes, habitually for some precise purpose. Ideally, a category illuminates an association among the subjects and objects of knowledge. Categorization is vital in language, prediction, inference, decision making and in all types of ecological communication. It is specified that categorization plays a key part in computer programming.Web mapping is the procedure of using maps brought by geographical information systems (GIS). Since a web map on the World Wide Web is together served and consumed, it is more than just web cartography; it is both a service action and consumer action. Web GIS stresses geodata processing traits extra involved with design traits such as data acquisition and server architecture such as algorithms and data storage, rather than it does the end-user reports themselves. The words web GIS and web mapping keep on slightly identical. Web GIS routines web maps and end users who are web mapping are acquiring investigative competences. The word location-based services denotes to web mapping user goods and services. Web mapping regularly includes a web browser or other consumer agent capable of client-server interactions.

From fig.1, it can be explained that, in this search engine user enters into the system first and login to use the categorized search and map. After user log into the system, the user types the query to carry out a search. The query undergoes three processes in this step. They are:

- Preprocessing
- Stemming
- Indexing

In preprocessing the data is converted into the form that the system can understand. It is also called cleaning of data. Then the processed data is stemmed. Stemming is the process of removing the suffix words such as s, es etc…Then the indexing process is done to build the list of words. Then the architecture is divided

into two branches one is for categorized result and the other for map. In first work, it uses Naïve Bayes theorem, it converts the data into table. Then likelihood is calculated and theorem is applied for classifying the results. The other work map search is carried in a separate page. When the user clicks the map button, it will lead to a new window. In that tab, user can search the place by typing the query; the web map for the query will be displayed in the screen. And additionally website of that place can also be opened. For example, if you search for Kalanikethan, then the website of that shop will be opened. This is overall search process of this search engine.

Algorithm Used: This system consists of following algorithm. It is given as follows:

- Naïve Bayes Algorithm

Steps for Searching Process:

- Preprocess the query.
- Remove the root words by applying the stemming algorithms.
- Build index using the indexing process as depicted in fig.2 to create a list of words where the results are present.

Naïve Bayes Algorithm: Naïve Bayes is one of the popular data mining algorithms which find its applications in Machine learning; Knowledge based system and medical diagnosis etc…The data for naïve Bayes need to be preprocessed for categorization. That is the format is converted into vector space notation [5]. The prosperity of Naïve Bayes depends on high rate of feature dependencies [6]. There are two models for text classification of Naïve Bayes. In One model, document is represented in the form vector of binary attributes. The other one represents document using word occurrences. Categorization is performed by calculating the posterior probability [7].

Algorithm Steps:

- Creation of table- Convert the given dataset into frequency table [8].
- Creation of likelihood table –The table is created by finding the probabilities, for example. Overcast probability [9].
- Applying Naïve Bayes theorem,

Where,
p(a|b) = Posterior Probability
p(b|a) = Likelihood
p(b) = Predictor probability
p(a) = Class prior Probability[10][11].

Geocoding: Geocoding is the process which is carried out to convert the query given by the user in to geographical coordinates such as latitudes and longitudes.Then the coordinates will be displayed on the web map using the process of address lookup. It is also called as Reverse Geocoding [12].

Experimental Results: The Search Engine is deployed in online with the following features: Categorized results and map search. Initially 20 categories were added, in each category 30-40 datasets are loaded in server. The results are examined by executing the search for different types of categories and analyzing it in terms of relevancy between the results, quality, speed etc… The second feature map is implemented by uploading the data in web server. The results are verified by searching any geographic location such as India gate etc. The geographic map will be displayed in the screen.

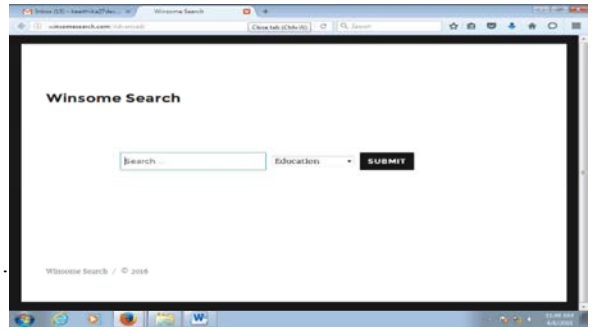

Fig.6 Home Page

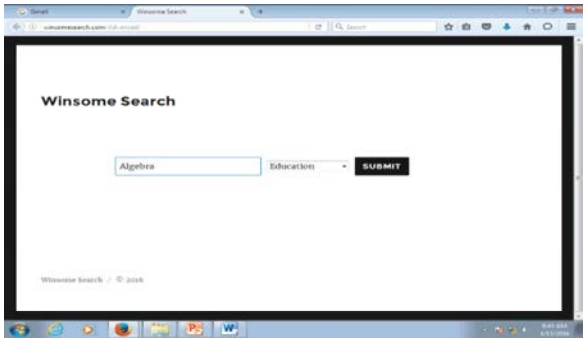

Fig. 7: Querying in the Search Engine

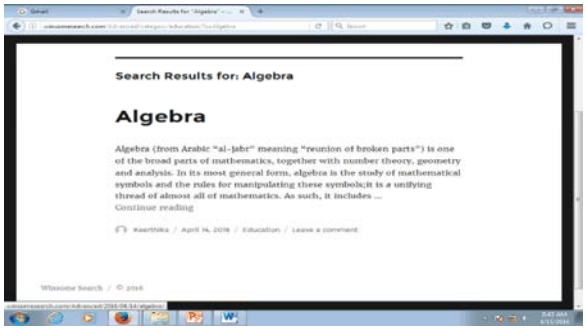

Fig. 8: Search results for the search engine

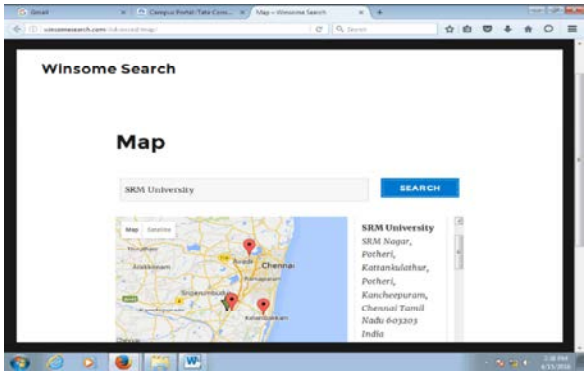

Fig. 9: Map Search

CONCLUSION

In this paper, we have designed a search engine with two features such as Categorization and Map search using Naïve Bayes Algorithm. Our empirical results shown that the results were provided in a categorized manner and relevant. Naïve Bayes is the main reason for the effective categorized results. The map results are accurate and user friendly. Our future work is to provide image and video search.

REFERENCES

1. Keerthika, D. and G. Sangeetha, 2015. A Survey on Web Search Engines, International Journal of Innovative Research in Computer and Communication Engineering, 3(10): 9233-9240.
2. Gautam Pant, Padmini Srinivasan and Filippo Menczer, 2004. Crawling the Web, Springer Web dynamics Adapting change in content size, topology and use, pp: 154-177.
3. Mini Singh Ahuja, Dr Jatinder Singh Bal and Varnica, 2014. Web Crawler: Extracting the Web Data, International Journal of Computer Trends and Technology (IJCTT). 13(3): 132-137.
4. Sergey Brin and Lawrence Page, 1998. The Anatomy of a Large-Scale Hyper textual Web Search Engine.
5. Susana Eyheramendy, David D.Lewis and David Madigan, 0000. On the Naïve Bayes model for categorization.
6. Rish, I., 0000. An empirical study of the naive Bayes classifier, pp: 41-46.
7. Andrew McCallum Kamal Nigam, 0000. A Comparison of Event Models for Naive Bayes Text Classification.
8. Eibe Frank and Remco R. Bouckaert, 0000. Naive Bayes for Text Classification with Unbalanced Classes.
9. Hiroshi Shimodaira, 2015. Text Classification using Naive Bayes,
10. Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim and Sung Hyon Myaeng, 2006. Some Effective Techniques for Naive BayesText Classification, IEEE Transactions on Knowledge And Data Engineering, 18(11): 1457-1466.
11. Eunseog Youn a and Myong K. Jeong, 2009. Class dependent feature scaling method using naive Bayes classifier for text data mining", Elsevier Pattern Recognition Letters, 30: 477-485.
12. Soumya Mazumdar, Gerard Rushton, Brian J Smith, Dale L Zimmerman and Kelley J. Donham, 2008. Geocoding Accuracy and The Recovery of Relationships Between Environmental Exposures and Health, Int J Health Geogr., pp: 7-13.