International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 12, Issue 1 (January 2016), PP.80-84

Data Extraction Using Search Engine for Exact and Accurate Information

Binoj M[1], Sreeja S P[2], Jincy Mathew[3]
[1] New Horizon College of Engineering ,Faculty, Bangalore, Karnataka, India.
[2] New Horizon College of Engineering, Faculty, Bangalore, Karnataka, India.
[3] New Horizon College of Engineering, Faculty, Bangalore, Karnataka, India.

Abstract:- Data Extraction (DE) is the function of extracting the essential knowledge from text automatically via search engines. The enormous volume of text presents now available on the World Wide Web a unique opportunity for DE. Data Extraction systems promise to encode vast quantities of Web content into machine processable knowledge bases, presenting a new approach to a fundamental challenge for artificial intelligence: the automatic acquirement of enormous volume of knowledge. Such knowledge bases would significantly extend the capabilities of Web applications. Future Web search engines, for example, could query the knowledge bases to answer complex questions that require synthesizing data across multiple Web pages. However, DE on the Web is challenging due to the enormous variety of diverse concepts expressed. All the data extraction techniques make errors. Standard error-detection strategies are used in previous and small-corpus extraction systems is intractable on the Web.

The increasing penetration of the real world with embedded and globally networked sensors leads to the formation of the Internet of Things, offering global online access to the current state of the real world. We argue that on top of this realtime data, a Web of Things is needed, a software infrastructure that allows the construction of applications involving sensor equipped real-world entities living in the Internet of Things. A Key service of present site is that getting so many information, without getting exact information. In this paper, we show how the existing Web infrastructure can be leveraged to get exact or accurate information without giving unnecessary details**.**

Keywords:- Precise data, Search Engine, relevant Information, Web page ranking, web crawler

I. INTRODUCTION

Wireless networks are at the epicenter of this trend. At its broadest, a wireless network refers to any network not connected by cables, which is what enables the desired convenience and mobility for the user. Not surprisingly, given the myriad different use cases and applications, we should also expect to see dozens of different wireless technologies to meet the needs, each with its own performance characteristics and each optimized for a specific task and context. Today, we already have over a dozen widespread wireless technologies in use: WiFi, Bluetooth, ZigBee, NFC, WiMAX, LTE, HSPA, EV-DO, earlier 3G standards, satellite services, and more.

As such, given the diversity, it is not wise to make sweeping generalizations about performance of wireless networks. However, the good news is that most wireless technologies operate on common principles, have common trade-offs, and are subject to common performance criteria and constraints. Once we uncover and understand these fundamental principles of wireless performance, most of the other pieces will begin to automatically fall into place. Wireless technologies can be classified in different ways depending on their range. Each wireless technology is designed to serve a specific usage segment. The requirements for each usage segment are based on a variety of variables, including Bandwidth needs, Distance needs and Power.

Wireless Wide Area Network (WWAN): the Internet via a wireless wide area network (WWAN) access card and a PDA or laptop. These networks provide a very fast data speed compared with the data rates of mobile telecommunications technology, and their range is also extensive. Cellular and mobile networks based on CDMA and GSM are good examples of WWAN. [6] Wireless Personal Area Network (WPAN): These networks are very similar to WWAN except their range is very limited. The main purpose of a WSN is to provide users with access to the information of interest from data collected by spatially distributed sensors. In real-world applications, sensors are often deployed in high numbers to ensure a full exposure of the monitored physical environment. Consequently, such networks are expected to generate enormous amount of data. The

desire to locate and obtain information makes the success of WSNs applications, largely, determined by the accuracy and quality of the extracted information.[2]

The principal concerns hen extracting information include the timeliness, accuracy, cost, and reliability of the extracted information and the methods used for extraction. The process of Data Extraction enables unstructured data to be retrieved and filtered from sensor nodes using sophisticated techniques to discover specific patterns. Practical constraints on sensor node implementation such as power consumption (battery limits), computational capability, and maximum memory storage, make DE a challenging distributed processing task. In terms of data delivery required by an application, DE in WSNs can be classified into four broad categories: eventdriven, time-driven, query-based, and hybrid. In event-driven, data is only generated when an event of interest occurs, while, in the time-driven, data is periodically sent to a sink every constant interval of time. With query-based, the data is collected according to end user's demand.[8] The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained indices with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Search Engine is giving so much information without giving accurate information. By using this Search engine the user or customer is able to get accurate information and in precise form.

II. RESEARCH GAP

The web accords a miraculous opportunity and challenge for web mining because of its following nature:

1. The information available on the web is huge, wide, diverse and dynamic.
2. It contains a mix of data such as text, video, audio, image etc.
3. Information on the web is heterogeneous, unstructured, semi structured and hyper linked.
4. A lot of information on the web is redundant and noisy.
5. The web consists of surface web and deep web.
6. Furthermore, web is also about services and it forms a virtual society.

Due to these characteristics, web mining has received much attention among many researchers. Some of the research areas in web mining are : structuralizing the web, vertical search, mobile search, multimedia search, object-level search, deep web search, link analysis and web data mining, learning to rank and find search relevance, document information extraction, removal of duplicates and improving the performance of search engines, tracking user behavior and predicting trends to promote business

III. LITERATURE SURVEY

For this proposed research the following literatures are reviewed

In News Keyword Extraction for Topic Tracking by Sungjick Lee ,[2008] mentioned[8] news keyword extraction from wide topic. Here they are extracting keyword from the news for easy understanding. But no extraction through search engine and no merging of search engine

In online discovery of relevant terms from internet Ji Donghong, Yang Lingpeng, Nie Yu, Tang Li , [4] the relevant keyword is first found and matching with contents , instead of searching the details summary values from net. But here it is not making common values by using different search engine.

In An Optimized Key Frame Extraction for Detection of Near Duplicates in Content Based Video Retrieval Sudeep D. Thepade , Ashvini A. Tonge, [2013] mainly focused for finding out duplication of videos which is uploading. Here not giving much importance to text.

In Visual Re-rank: An Approach for Image Retrieval from Large-scale Image Database Pushpanjali M. Chouragade.[3] , removing unrelated image by ranking and put proper image. Normally in yahoo and google importance is giving text related image. But in it is not giving importance to text extraction

In A Generic Framework for Video Search Using Feature Extraction and Annotation Sunil Parihar , Kshitij Patha,[2008] it is going[8] to make clarity for top ranking video. But here not mensioning the importance to text extraction.

Data Utku Irmak, Vadim von Brzeski, Reiner Kraft,[2009] the keywords are checking with number of sites visited. The predefined keywords are matching with present keywords in the website. So, this both click will be equal to how many times the website visited. But here it is not mensioning searching of whole data together.

T.Seeniselvi, R.Imrankhan, [2013]conducted research with the name Personalized Mobile Search Engine by Analyzing Query Travel Patterns with Association Rule Mining. Another research conducted on Efficient Search in Large Textual Collections with Redundancy. Here repeated archive file is compressing through crawler .

Sergey Brin and Lawrence Page presented the search engine working like text acquisition, text transformation and through index creation then index. From text acquisition, it is going to document data store also. Text transformation means transforming documents into index terms or features. Index terms: the parts of a document that are stored in the index and used in searching. Text transformation means, Processing the sequence of text tokens in the document to recognize structural elements such as titles, figures, links, and headings. The main key sources of research topic are that redundancy, information retrieval, time complexity and cost optimization. The main problem of here is, when people search on different networks or acquiring the information needed, it may arise confusion in selecting the search engine.

Regarding text mining, by Ian H.Witten, [2] it merely depends upon how to access a particular text from whole text. Here it is not mensioning how we can access common data or similar data in every search engine. Data extraction for search engine results pages using visual cue and DOM tree by Jer hang cong is doing about searching the content in different way by using tree for getting correct data. Here they mensioned how data is going to search and not extracting correct or exact data.

In A Survey of Text Mining Techniques and Applications by Vishal Gupta summarizes about text mining or text selection fast compared to normal web search. Here also not mentioning how will access the accurate data.

In Automatic Extraction from deep web by Nripendra Narayan Das, Ela Kumar about selection of unstructured data properly. Here not mentioning about text extraction as a whole data.

According to Content Based Image Searching Using Focused Crawler by Ayush Agrawal, Kanchan Hans [15] age searching. It involves text extraction from images and use the crawling approach through extracted text.

According to A Scalable Image Snippet Extraction Framework for Integration with Search Engines by Sheikh Muhammad Sarwar, Md. Mustafizur Rahman [16] by we propose and implement a search framework by integrating text and image search engines that increases the speed of extracting a representative image of a web document.

According to Implementation of a Web Search Engine for Restaurants using Lucene1 by Sojourn Oh, Minsoo Leesxz, if a user wants to know more information about the restaurant, the user can connect to the website.

Regarding Intelligent Semantic Web Search Engines: A Brief Survey by G. Madhu and Dr. A. Govardhan, search engines used to retrieve relevant and meaningful information intelligently.

IV. PROBLEM STATEMENT

The existing approaches are inadequate in providing a better solution for the people when they search on different networks for essential information needed for making decisions. So it may lead to confusion on selecting the search engines. In certain situations the search engine also fails to provide the correct results as per the perception.

- Redundancy**:** The redundancy or duplication of data is going to happen when we are searching the data from search engines because of limited data. New decisions and innovations can be made in time as per the needs of the consumers which leads to the company's profitability. With the extracted information, the market research people can make new products as per the perceptions of the customer within stipulated time if correct data is there.
- Time Complexity**:** In current approaches it is very difficult to find out exact data within a specified time. So, it is extremely hard to find out the necessary information within a specified time because the people are confused about the selection of search engine also. So, the company is not able to find out how much time is spent for the specified, if and only if the common data is not identified.
- Cost Optimization: The proposed approach can be used by the companies for acquiring the essential data in time. It is an advantage for the companies for making innovations in time leads to cost optimization. This timely data enables the company to greater heights in terms of profit and competitiveness.

V. OBJECTIVES OF THE RESEARCH

The proposed research is primarily devised to achieve the following objectives:

1. To extract data from search engines using wireless networks without any duplication
2. To design an approach for data extraction, data in time
3. To develop a cost effective approach for extracting unique data from the web
4. To investigate and manage a database of urls and the keywords
5. To design an approach to detect the redundancy in data extraction

VI. PROPOSED ARCHITECTURE

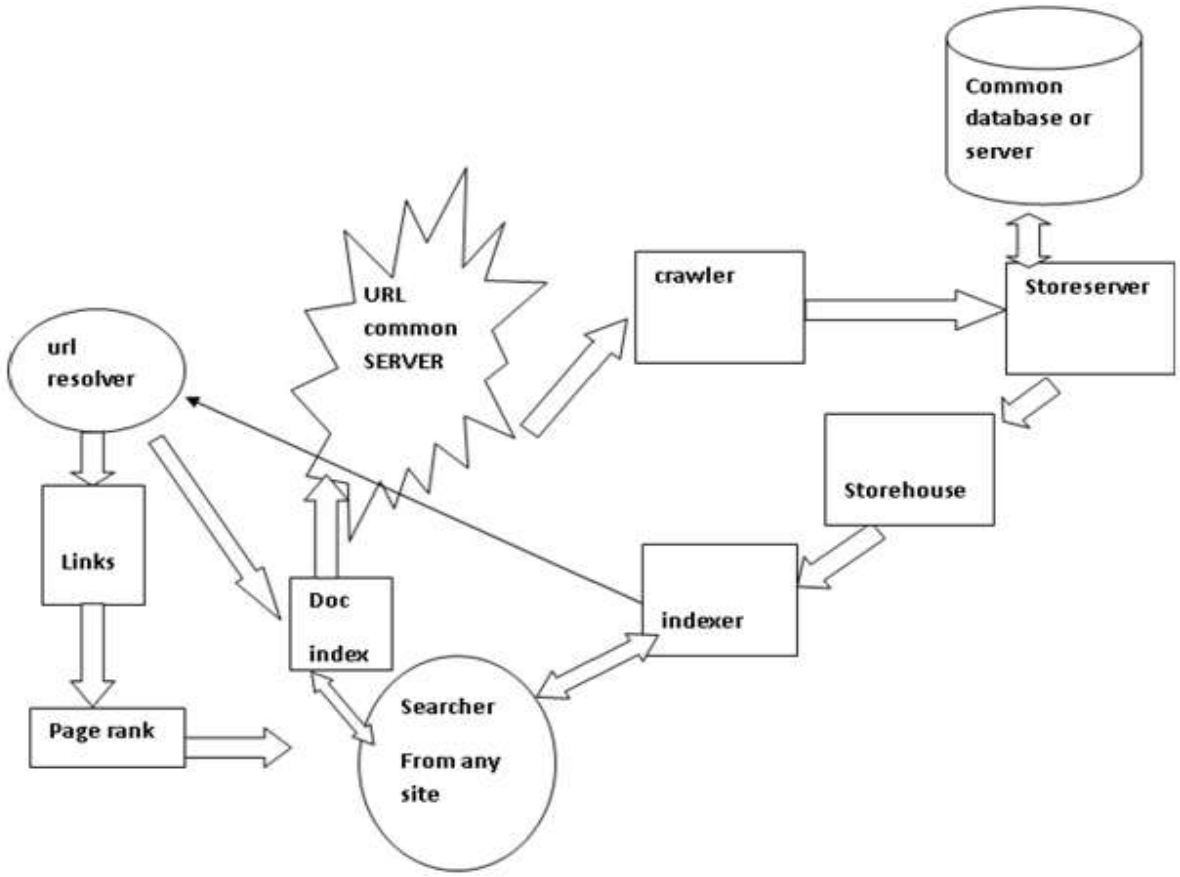

Fig. 1: Search engine architecture

Compared to Existing system proposed system, this is going to operate through an Index object. This object lets us store information via an index document, retrieve documents using search queries, modify documents, and edit documents. Each index has an index name. We can use namespaces and indexes to organize your text files. In our code we produce an Index object by setting the index name.

Documents hold an index's searchable content. A text file is a container for structuring index able data. From a technological point of view, a Document object represents a uniquely identified collection of disciplines, identified by a document ID. Fields are named, typed values. Text files do not have kinds in the same sense as Data store entities. Each document field has a unique field type. The type can be any of the following varchar: A plain text string in the module Search. Query a document's contents, we must add the text file to an indicator. Indexing allows the document to be searched with the Search API's query language and query option. Winning over a variety of formats (e.g., HTML, XML, PDF,…) into a consistent text and metadata format.

Simple database to handle large numbers of documents and structured information. Document components are typically stored in a compressed form for efficiency

Translating documents into index terms or features is called Text transformation.

The parts of a document that are stored in the index and used in searching. Roles of a text document that is utilized to interpret its content are called features. The circle of all the conditions that are indexed for a document collection is called Index vocabulary. Parsers will process the sequence of text tokens in the text file to recognize structural elements such as titles, shapes, links, and headings and Tokenization is the procedure of identifying units to be indexed. Then It will remove common words from the stream of tokens such as the, of, to etc. to reduce index size considerably.

VII. METHODOLOGY

The method following is by the means of indexing mechanism. Through various search engine users are exploring the different contents, which is moving to the common search engine. For this proposed system, the technology requirement is below. Java, jsp, servlet as front end and mysql as backend. In the database, the common database using as Mysql server. Tools required is eclipse.

This follows the typical search engine architecture discussed The main technologies used include HTML, Java, Java Server Page (JSP), Java Bean and Java Servlet. The architecture is shown below. The right-

hand side of the figure shows the batch process of building the search engine. First, a simple Vertical Spider developed in Java collects Web pages in the technology domain and calculates the number of inlinks (hyperlinks pointing to a page from other Web sites).

VIII. RESEARCH BACKGROUND/MOTIVATION

My closed topic is related to the current market trend. So, currently lot of people who are using depending on the search engine for so much information. Normally I am very much interested to develop web related projects and I am trying to do so much project my own by using web related technologies. I am so much interested to do web based project my own if I am getting time. In the present research, nobody trying to do to make a common search engine to do searching activities and no researcher try to do this much complex project. All research is related to fast extracting or extracting unstructured information.

IX. SCOPE OF THE PROJECT

Here mainly concentrating on the problem of people's confusion. Because the people are not in a position to get exact information what they want and the people are in so much confusion which search engine they have to depend for the information. The second problem is that the people's proper utilization of time for the productive work, because the people are lacking time for unproductive searching.

X. CONCLUSION

Search Engine is a widespread search engine that affects millions of people worldwide every year. Due to the alarming rate of the spread of information and knowledge, particularly in whole countries, IT professionals, teachers are implementing new strategies for the getting correct information. It is advisable to do search engine again fast with structured and correct information within stipulated time.

REFERENCES

[1]. Dr. A. Muthu Kumaravel1**, "**Mining User Profile Using Clustering From Search Engine Logs", Vol. 2, Issue 6, June 2014

[2]. Ian H. Witten , "Text mining",

[3]. Pushpanjali M. Chouragade , Prashant N. Chatur "Visual Re-rank: An Approach for Image Retrieval from Large-scale Image Database", Volume 3, Issue 3, March 2013.

[4]. Ji Donghong, Yang Lingpeng, Nie Yu, Tang L,i ,"ONLINE DISCOVERY OF RELEVANT TERMS FROM INTERNET"

[5]. Utku Irmak, Vadim von Brzeski, Reiner Kraft, Contextual Ranking of Keywords Using Click Data ,2009.

[6]. T.Seeniselvi1, R.Imrankhan2"**,**Personalized Mobile Search Engine by Analyzing Query Travel Patterns with Association Rule Mining", Volume 2, No.9, August 2013

[7]. Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine"

[8]. Sungjick Lee, "News Keyword Extraction for Topic Tracking" 2008

[9]. Sunil Parihar , Kshitij Pathak, A Generic Framework for Video Search Using Feature Extraction and Annotation , Volume 3, Issue 6, June 2013

[10]. V. Jayashree, Mr. A. Balasubramanian, Web Search Recommendation System Using Concept Based Mining Techniques, Volume 4, Issue 7, July 2014

[11]. Pinky Paul, ENTITY SEARCH ENGINES, Vol.3 Issue.2, February- 2014, pg. 877-880

[12]. P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, Content Based Ranking for Search Engines, vol 1, March 14-16 2012.

[13]. Nripendra Narayan Das, Ela Kumar automatic extraction of data from deep web

[14]. page, ISSN 2347 – 8527 Volume 3, Issue 1 April 2014

[15]. Michael J. Cafarella, Oren Etzioni, Dan Suciu , Structured Queries Over Web Text, 2011

[16]. Ayush Agrawal , Kanchan Hans, Content Based Image Searching Using Focused Crawler, Volume 1, Issue 1,Pages-13-17, November -2014.

[17]. Sheikh Muhammad Sarwar, Md. Mustafizur Rahman, Md. Haider Ali, Ashique Mahmood AdnanA Scalable Image Snippet Extraction Framework for Integration with Search Engines, Vol. 6, No. 1; 2013 ISSN 1913-8989

[18]. Sojourn Oh , Minsoo Lee, Computer Science and Engineering, Implementation of a Web Search Engine for Restaurants using Lucene1, ISSN 2320 – 2602 Volume 4 No.4, April 2015

[19]. G. Madhu and Dr. A. Govardhan, Dr. T. V. Rajinikanth, Intelligent Semantic Web Search Engines: A Brief Survey, (IJWesT) Vol.2, No.1, January 2011.