International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

An Effective Framework for Cloud Based Search Engine

Dr. R. Malathi Ravindran
Associate Professor, Computer Applications, NGM College, Pollachi, Tamil Nadu, India

ABSTRACT

In present world, we have plenty of information around us. However, for a specific information need, only a small subset of all the available information will be useful. Over the 1970's and 1980's, much of the research in information retrieval was focused on document retrieval, and the emphasis on this task in the Text Retrieval Conference (TREC) evaluations of the 1990's has further reinforced the view that information retrieval is synonymous with document retrieval. Web search engines are, of course, the most common example of this type of information retrieval system. The enormous increase in the amount of online text available and the demand for access to different types of information have, however, led to a renewed interest in a broad range of information retrieval related areas that go beyond simple document retrieval, such as question answering, topic detection and tracking, summarization, multimedia retrieval (e.g., image, video and music), software engineering, chemical and biological informatics, text structuring, text mining and genomics. In this paper, Cloud Computing Based Information Retrieval (CCBIR) system is introduced for the information retrieval from the huge volume of data.

Keywords : Search Engine, Cloud Computing, Information Retrieval, Document and CCBIR.

I. INTRODUCTION

Search Engines has become an essential need for internet users in their today's day to day life. In the present world, top Search engines are earning profit from advertising, entertainment, social media networks, daily use applications (e.g. maps) and online product sales and services. Advertising through Internet about their online business is increasing day by day. It shows the tremendous influence of Internet. In this scenario, the need of search engines for the users is very much essential. There are lot of search engines available, including web search engines, selection-based search engines, meta search engines, desktop search tools, web portals and vertical market websites that have a search facility for online databases. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Every search engine has its own merits and demerits. The proposed system in this work called Cloud Computing Based Information Retrieval (CCBIR) is developed for the efficient information retrieval [4].

II. SEARCH ENGINES

A. Existing Search Engine Models

Privacy Search Engines

Two types of privacy search engines are available such as DuckDuckGo and Ixquick. DuckDuckGo (DDG) is an internet search engine that emphasizes protecting searchers' privacy and avoiding the filter bubble of personalized search results [8]. DuckDuckGo distinguishes itself from other search engines by not profiling its users and by deliberately showing all users the same search results for a given search term[13]. Ixquick (styled "ixquick") is a meta search engine based in New York and the Netherlands [www.metamend.com] founded by David Bodnick in 1998. Ixquick is owned by Surfboard Holding BV of the Netherlands, which acquired the internet company in 2000. Ixquick provides the stand-alone proxy service[6].

Open Source Search Engines

DataparkSearch, Gigablast, Grub, ht://Dig, Isearch, Lemur, Toolkit & Indri Search Engine, Lucene, mnoGoSearch, Namazu, Nutch, Recoll, Sciencenet (for scientific knowledge, based on YaCy technology), Searchdaimon, Seeks, Sphinx, SWISH-E, Terrier Search Engine, Xapian, YaCy and Zettair are some of the open source search engines. DataparkSearch is a search engine designed to organize search within a website, group of websites, intranet or local system. DataparkSearch is written in C [7]. Grub is an open source distributed search crawler platform. Users of Grub can download the peer-to-peer grub client software and let it run during computer idle time. ht://Dig is a free software indexing and searching system created in 1995 by Andrew Scherpbier [12]. Lemur toolkit is used for developing search engines, text analysis tools, browser toolbars and data resources in the area of IR. Apache Lucene is a free open source information retrieval software library, originally written in Java by Doug Cutting. mnoGoSearch is an open source search engine for Unix-like computer systems written in C. Searchdaimon ES (Enterprise Search) is an open source enterprise search engine for full text search of structured and unstructured data available under the GPL v2 license. SWISH-E stands for Simple Web Indexing System for Humans - Enhanced. It is used to index collections of documents ranging up to one million documents in size and includes import filters for many document types. Xapian is an open source probabilistic information retrieval library, released under the GNU General Public License (GPL). It is a full text search engine library for programmers. It is written in C++, with bindings to allow use from Perl, Python, PHP, Java, Tcl, C#, Ruby and Lua [6].

Semantic Browsing Engines

There are two types of semantic browsing engines in use such as Hakia and Yebol. Hakia was founded by Rıza Can Berkan, a nuclear scientist by training with a specialization in artificial intelligence and fuzzy logic, and PenttiKouri, a New York-based economist and venture capitalist[11]. Yebol was a vertical "decision" search engine that had developed a knowledge-based, semantic search platform [6].

Social Search Engines

ChaCha Search, Delver, Eurekster, Mahalo.com, Rollyo, SearchTeam, Sproose and Trexy are some of the popular social search engines. ChaCha is a human-guided search engine. It provides free, real-time answers to any question through its website or by using one of the company's mobile apps. Eurekster is a company founded in Christchurch, New Zealand, that builds social search engines for use on websites. Trexy is an internet meta search engine established in early 2006. It allows users to record and share "search trails" of their activity on search engines [6].

Visual Search Engines

ChunkIt, Grokker, Pixsta, PubGene, TinEye, Viewzi and Macroglossa are some of visual search engines. A tech company Groxis developed and marketed the web-based content access and visual search engine called Grokker. Pixsta Ltd. is a UK-based image and video search company founded in 2006 by Alexander Straub, Dr. Daniel Heesch and David Williams. PubGene AS is a Bio informatics company. In 2001, PubGene founders demonstrated one of the first applications of text mining to research in biomedicine (i.e., biomedical text mining). Macroglossa is a visual search engine based on the comparison of images, coming from an Italian Group. Users can upload photos or images that they aren't sure what the images contain [6].

Search Appliances

Google Search Appliance, Fabasoft, Munax, Searchdaimon and Thunderstone are some of the available search appliances. The Google Search Appliance is a rack-mounted device providing document indexing functionality that can be integrated into an intranet, document management system or web site using a Google Search-like interface for end-user retrieval of results. The operating system is based on CentOS. The software is produced by Google and the hardware is Thunderstone, a US-based software company specializing in enterprise search [14]. Fabasoft AG is a software manufacturer headquartered in Linz, Upper Austria. The company was established in 1988 by Helmut Fallmann and Leopold Bauernfeind [6].

Desktop Search Engines

Copernic Desktop Search, Everything, Google Desktop, InSight Desktop Search and Tropes Zoom are some of the popular available desktop search engines. Usenet-Google Groups is a service from Google Inc. that provides discussion groups for people sharing common interests. Google Groups offers at least two kinds of discussion group; in both cases users can participate in threaded conversations, either through a web interface or by e-mail. Through the Google Groups user interface, users can read and post to Usenet groups [Chuq Von Rospach (1999)] [6].

B. Drawbacks of Existing Search Engines Model

- **Poor precision** - List of retrieved documents contains a high percentage of irrelevant documents.
- **Poor recall** - Most web's search engines consult databases of the most frequently used words in documents, such as words drawn from documents title and first few sentences. Hence the search engines won't retrieve documents in which the keywords for which the user is searching are buried somewhere within document. Many page authors send search engine, a numerous web pages containing various tricks like irrelevant title tag or repeating certain words in first few levels that are irrelevant to actual contents of the page to boost the ratings. Though this seems to be matter of less concern but when attempted by many people leads to very serious problems. It might lead to situation where in not even one of the top ten sites listed would be of subject expected.
- **Varied document quality** - spider can't discriminate between valuable documents and spams.
- **Varied indexing depth** - Some spiders retrieve only the document's title, others retrieve entire document text. Unless the user understands the spider's work, he is not very likely to succeed.

III. INFORMATION RETRIEVAL

A. Challenges in Information Retrieval

As a search engine, the user is damned and heavily criticised by all and sundry. They have to face a lot of challenges while retrieving the relevant information from the huge volume of data [3].

- **Discovery** - The first challenge for any search engine is to actually find out the content there in the first instance. In some cases it will be easy because there are sites that use Ping O Matic which alerts search engines every time when a new content is published or a new website is born. In some cases it is even easier, when an existing and trusted resource links to a web page that the search engine didn't previously know about. In the absence of both, other options may include the webmaster informing the search engines via a submission form. The additional challenges presented are pages that have changed in some way be it a new URL, updated content, updated formatting i.e. how the content is presented to the user which may include technologies, media formats and page layouts.
- **Storage** - Once it is determined that there is new content, the search engines then have the unenviable task of storing copies of the content found on the internet. There are over a several trillion known web pages let alone a googal of unknown published web pages on the net. The computer resource to construct the storage facility for this volume of information is quite unthinkable.
- **Extraction** - Not all of the data found on websites are actionable for providing a service for search engine users. So search engines are likely to extract certain elements or attributes of the web page in order to get some sense of a web page's makeup. This will not only help with modelling but also with the recall of information.
- **Modelling** - The modelling is what transforms the stored data into information and that means using (although not necessarily) machine learning to detect patterns that helps search engines determine sites on the basis of authority, relevancy, user experience and readability. This would require dictionaries on world languages, search query logs, user behaviour logs on search engines, user experience technologies to say the least.
- **Recall** - Once modelled, the search engines have to produce the results of how they ordered and organised the information, in a timely fashion. This is resource intensive if the search engine is well known and widely used across different languages, as users are likely to be incredibly demanding and will want to know the results near instantly. They will also expect the results to be satisfactory such that no further searches will need to be performed in order to find what they are looking for. One of the ways search engines will determine this will be whether the document leads to a new type of follow up search query or not. For example, now that the searcher read the brand comparison site page, and the searcher is searching

on brand, the searcher can conclude the site page is relevant to the previous search result.

B. Security Aspects in Information Retrieval

Security of information majorly designs to protect the three parameters of the C.I.A i.e. Confidentiality, Integrity and Availability. Now a day's huge amount of work is being done to protect the information security of the organization. It's a privilege based procedure to protect the company's assets and resources from unauthorized disclosure. Fraudulent activities are executed in the banking industry and other sectors of public interference. In the recent times, banking industry is not able to stop the fraud before it happens.

According to Guardian Analytics, Banking industry is not able to probe the 78% of the online fraud where the attacker intentionally retrieves the information online like account number, name and pin. Thus misuses for one's one end. The organizations generally do not possess the required tools to protect the attacks which are implanted with negative intention. The organizations are installing firewall, anti viruses and UTM devices. Majorly they are following the proactive approach in order to protect from malicious activity with a backup mechanism [1].

IV. CLOUD COMPUTING

Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources. It is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models [9] [10].

A. Basic Concepts, Goals and Benefits

There are a set of basic terms that represent the fundamental concepts and aspects pertaining to the notion of a cloud and its most primitive artifacts.

Cloud - A cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources. The term originated as a metaphor for the Internet which is, in essence, a network of networks providing remote access to a set of decentralized IT resources. Prior to cloud computing becoming its own formalized IT industry segment, the symbol of a cloud was commonly used to represent the Internet in a variety of specifications and mainstream documentation of Web-based architectures [9]. This same symbol is now used to specifically represent the boundary of a cloud environment, as shown in Figure 1.

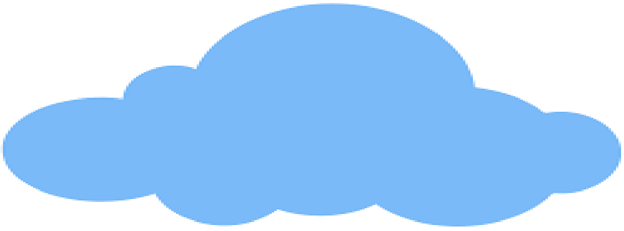

Figure 1. Cloud Symbol

The goals and benefits of the cloud computing are as follows;

- Reduced investments and proportional costs
- Increased scalability
- Increased availability and reliability

For a significant number of ICT decision makers today, cloud computing is a catalyst, if not a prerequisite, for innovation and transformation. The cloud approach is an important component in many e-governance, education and healthcare strategies worldwide [5][9].

Cloud computing can greatly benefit public sector organizations of all types and sizes by:

- Reducing costs and controlling costs: Consolidates facilities, optimizes human capital, utilizes assets efficiently, reduces capital expenditure and charges for services.
- Improving agility and adaptability: Virtualizes resources, increases capacity with simple scalability, expands or contracts services to meet demand, deploys software quickly, expands flexibly to meet needs.
- Enhancing services and collaboration: Takes advantage of leading-edge applications provides broad access for stakeholders and improves collaboration.

- Addressing risk issues: Maintains critical service levels, helps ensure resilience, and chooses cloud computing options that meet security and privacy requirements.

B. Cloud Computing Trends

In January 2015, RightScale conducted its fourth annual State of the Cloud Survey of the latest cloud computing trends with a focus on infrastructure-as-a-service. The survey questioned 930 IT professionals about their adoption of cloud infrastructure and related technologies. The respondents ranged from technical executives to managers and practitioners and represented organizations of varying sizes across many industries. The margin of error was 3.2 percent. The following diagram shows the survey report [9] [15].

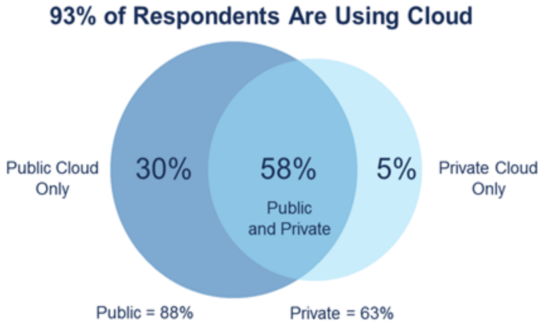

Figure 2. Cloud Computing Trends: 2015 State of the Cloud Survey

V. CLOUD COMPUTING ON INFORMATION RETRIEVAL

Cloud computing encompasses a pay-per-use paradigm for providing services over the internet in a scalable manner. Supporting data intensive applications is an essential requirement for the clouds. However, dynamic and distributed nature of cloud computing environments makes data management processes very complicated, especially in the case of real-time data processing/database updating [2][9].

A. Overall Information Retrieval System Flowchart

The working principle of overall system has been clearly represented in the following graph. In this system user submits a query to retrieve information. The CCBIR system retrieves the query from the user and collects the data from data source. After performing various processes internally, the system provides results to user [4]. Figure 3 clearly depicts the overall working principle of this process.

A. Overall Information Retrieval System Flowchart
The working principle of overall system has been clearly represented in the following graph. In this system user submits a query to retrieve information. The CCBIR system retrieves the query from the user and collects the data from data source. After performing various processes internally, the system provides results to user [4]. Figure 3 clearly depicts the overall working principle of this process.

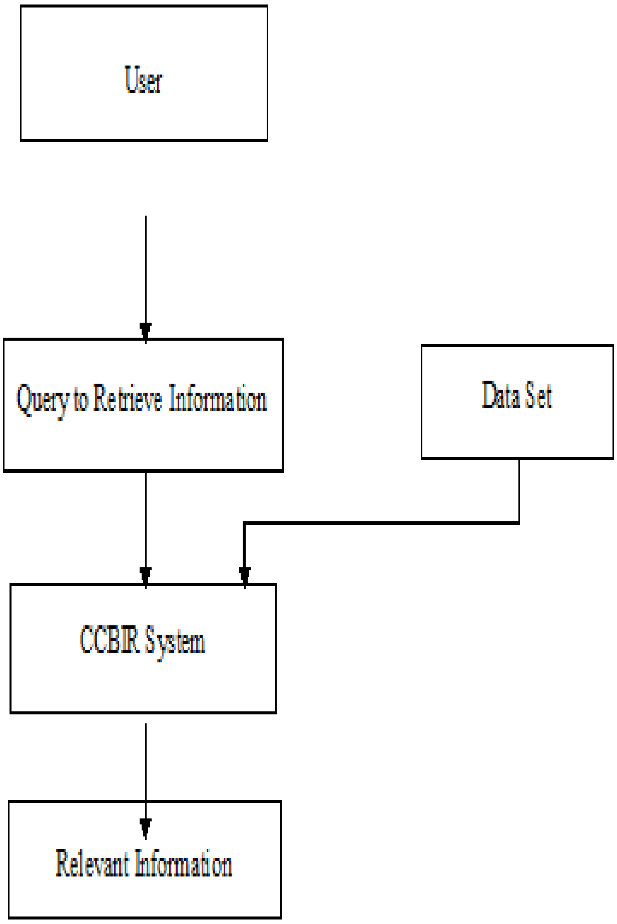

Figure 3 Overall Information Retrieval System Flowchart

B. CCBIR System Flowchart

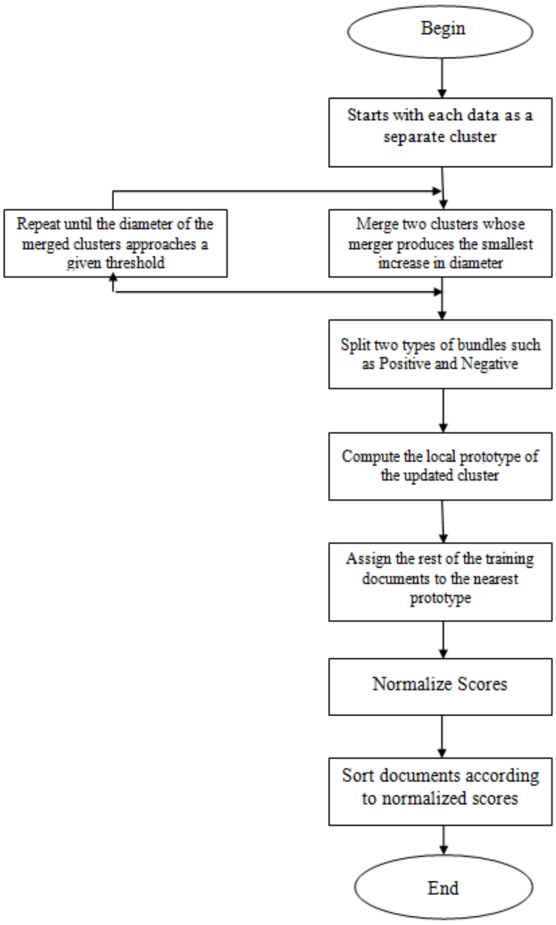
Begin
Starts with each data as a separate cluster
Merge two clusters whose merger produces the smallest increase in diameter
Repeat until the diameter of the merged clusters approaches a given threshold
Split two types of bundles such as Positive and Negative
Compute the local prototype of the updated cluster
Assign the rest of the training documents to the nearest prototype
Normalize Scores
Sort documents according to normalized scores
End

Figure 4. Working Principle of CCBIR System

VI. CONCLUSION

Information retrieval in web is different from retrieval in traditional indexed databases. The differences are due to the high degree of dynamism of the Web, it's hyper-linked character, the absence of a controlled indexing vocabulary, the heterogeneity of document types and authoring styles, the easy access that different types of users may have to it. Thus to assess the retrieval performance effectively using Web search engines, the new system called CCBIR is constructed using Vector Space Model based on Semi-Supervised Clustering. The experimental results of CCBIR system provides the better results than the existing information retrieval mechanism such as Meta, Boolean and Probabilistic.

VII. REFERENCES

[1]. Ashish Gautam et al.(2013), "Security Issues and Accuracy Concerns in the Information Retrieval Process", International Journal of Computer Applications, Vol. 70, No. 3, pp.1-6.

[2]. Amir H.Basirat and Asad I.Khan (2010), "Evolution of Information Retrieval in Cloud Computing by Redesigning Data Management Architecture from a Scalable Associative Computing Perspective", Springer Link, Neural Information Processing. Models and Applications Lecture Notes in Computer Science Vol. Part II, pp 275-282.

[3]. Andreas Voniatis, "Information Retrieval Challenges", www.alchemyviral.com/information-retrieval-challenges#.VTiJEPk70Zw, January 08, 2014.

[4]. Dr. R. MalathiRavindran (2017), “A Novel Approach for Information Retrieval Using CCBIR System”, IJSRCSEIT – International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol.2, Issue-1, ISSN: 2456 – 3307.

[5]. Fernando Macias and Greg Thomas, "Cloud Computing Advantages in the Public Sector", white paper, CISCO pp. 1-8.

[6]. HariPriyanka R, Malathi Ravindran R. (2016) “ A Cloud Computing with Search Engines”, IJSRD – International Journal for Scientific Research & Development, Vol.4, Issue 08, ISSN(online) 2321-0613.

[7]. Ixquick Q&A, Ixquick, January 2009. Retrieved 8 December 2013.

[8]. Jon Buys (2010). "DuckDuckGo: A New Search Engine Built from Open Source". GigaOM OStatic blog.

[9]. Malathi Ravindran R. and Antony Selvadoss Thanamani (2015), “A Novel Information Retrieval System for Effective Acquisition of Data using Cloud Computing”, International Journal for Science and Research in Technology , Vol.1, No. 8.

[10]. Thomas Erl, Zaigham Mahmood, and Ricardo Puttini (2013), "Cloud Computing Conepts, Technology & Architecture", Prentice Hall, ISBN-10: 0-13-338752-6.

[11]. Weisenthal, Joseph (2007), "Hakia Raises $2 Million for Semantic Search" (HTML), Hakia Raises $2 Million for Semantic Search, Retrieved 2007-12-03.

[12]. Winston A. (2003) "OpenVMS with Apache, OSU and WASD: The Nonstop Webserver", page 179, Digital Press.

[13]. www.dontbubble.us (2014)

[14]. www.news.cnet.com

[15]. www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2015-state-cloud-survey