AN EMERGING DATA MINING TECHNIQUE OVER CLOUD BASED ON METADATA INDEXING

Shital P.Shende[1], Prof.C.M.Mankar[2], Rupali P.Shende[3]

[1]*Master of Computer Engineering Department, S.S.G.M.C.E, Shegaon, (India)*
[2]*Assistant Professor Computer Engineering Department S.S.G.M.C.E, Shegaon, (India)*
[3]*Master of Computer Application, Saraswati College, Shegaon, (India)*

ABSTRACT

Research in peer-to-peer file sharing systems has targeted on try the planning constraints encountered in distributed systems, whereas very little attention has been dedicated to the user experience: these systems continually assume the user is aware of the concerning the file they're looking out. Nevertheless average users seldom even apprehend that file exists. File sharing systems that do take into account the user expertise and permit users to go looking for files by their name, usually gift centralized management and that they show many severe vulnerabilities, that build the system unreliable and insecure. the aim of this method is to style a additional complete distributed file sharing system that's not solely trustable, ascendable and secure, however additionally leverages the user's psychological feature employment. We tend to gift a technique that by mining a file's info designates relevant keywords for the file mechanically.

These keywords area unit later utilized for the file search and retrieval. Our system provides smart style principals from previous distributed file sharing systems to supply a trustable, scalable, secure and novel distributed file sharing system that a mean user might utilize for file search.

Search engines usually include a crawler that traverses the net retrieving documents and a hunt frontend that provides the program to the non inheritable info. The evolution of search engines nowadays is fast by provision additional search capabilities like a hunt for data moreover as search inside the content text. linguistics internet standards have provided strategies for augmenting files with data.

Keywords- Cloud Computing, Distributed System, Hygiene Factor , Lookup Problem, Search Engine.

I. INTRODUCTION

Basically evaluation of search engine is the process of making judgment about the value, importance and quality of search engine, after considering search engines carefully. The evaluation of search engines has not been keeping up with the advancement of their development. Web search engines work differently based on different mode of interface, features, coverage of the web, ranking methods, delivery of advertising and many more such factors. It is not easy to evaluate them on a single basis. There are many strategies for evaluating search engines such as automatic evaluation, human relevance judgment based evaluation. The purpose of this paper is to review the search engine evaluation strategies in order to propose an enhanced method for evaluating search engines. Distributed systems are a collection of autonomous computers connected through a network. A distributed system permits the computers to share resources and activities, allowing the end user to perceive the system as a powerful single computing machine. Peer-to-peer systems are a particular type of distributed systems, where all computers, also known as nodes, present identical responsibilities and capabilities. Peer-to-peer systems have many advantages over traditional centralized systems: they present better availability, scalability, fault tolerance, lower maintenance costs as well as lower operation and deployment costs. The drawback of these systems is that they encounter several design challenges. For example the system must remain functional, despite the varying number of uncontrolled participating nodes. Furthermore the system must be decentralized and symmetric; load should also be balanced among all nodes. Additionally, despite the system's size, data search on peer-to-peer systems must be fast and robust (scalable).

A vast number of researchers have concentrated on solving the design challenges referred above. A problem that has been widely tackled is the lookup problem. The lookup problem assumes that a node A inserts a file x into the system and moments after, a node B seeks to retrieve the file x. Considering that the node A is no longer online, the lookup problem intends to find the location of a node that has a replica of the file x. Examples of novel architecture algorithms that were proposed to solve the lookup problem. Systems that also solved the lookup problem, Because of the characteristics of these systems, it is possible to utilize them as a base for developing more complex distributed systems. PAST is a large scale internet based global storage utility that provided scalability, high availability and security. With PAST users were capable of inserting files into the system and later retrieving them, or retrieving files that other users shared. It is important to note, that to accomplish this operation, the user needed to know the file's public key. PAST looked up files by utilizing Pastry. PAST made several improvements to file sharing but because PAST's lookups were based on the file's public key, the system doubtlessly encountered many usability problems. In specific, new users that were unaware of the existence of the public keys would be incapable of finding their file of interest. To overcome this problem, a centralized web server, that provided the public keys to the files the users were searching for, would be required. But adding a centralized web server to the system would increase the system's vulnerabilities to single points of failure. Additionally PAST did not handle all of the design issues encountered in peer-to-peer systems. Specifically it did not address load balance: PAST made no partition on the files that were inserted. Therefore if a large file was attempted to be added to the system, if it did not fit in one single node, the file would not be inserted, despite the fact that the system as a whole might present sufficient memory.

Another interesting large scale peer-to-peer storage system was Pond an implementation of Ocean Store. Pond presented several improvements and differences over PAST, the only problem was that Pond presented the same usability issue PAST encountered: the system required the user to know the public key of the file they were searching. A file sharing system, which did consider in more detail the user experience when sharing and seeking files is Bit torrent. Bit torrent is a file downloading protocol that together with sites, such as Piratebay.org, Lokotorrent.com and trackers servers provides probably the biggest distributed file-sharing system in the world Web pages supporting Bit torrent function by showing for each available file, its name, size, current numbers of downloader's and seeds, and the name of the person who uploaded the file. To download the file a user clicks on a link those points to a .torrent meta-data file. The .torrent metadata files are stored and distribute among .torrent file servers. This mechanism, permits users to search for files by simply inputting related keywords of the file name and querying a web server. Albeit Bit torrent presented a significant improvement on user experience in file sharing systems, Bit torrent is not a truly distributed system.

II. DEVELOPMENT MODELS OF CLOUD

Cloud computing is the collection of virtualized and scalable resources, capable of hosting application and providing required services to the users with the "pay only for use" strategy where the users pay only for the number of service units they consume. It allows consumers and businesses to use applications without installation and access their personal files at any computer with internet Access. Cloud Computing has attracted the giant companies like Google, Microsoft, and Amazon and considered as a great influence in today's Information Technology industry.

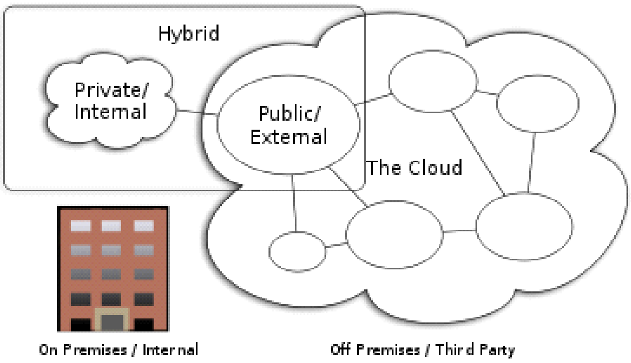

Fig.1 Types of Cloud

2.1 Public Cloud

This infrastructure is available for public use alternatively for a large industry entity and is closely-held by an organization selling cloud services. A Public cloud is one based on the standard cloud computing model, in which a service provider makes resources, such as applications and storage, available to the general public over the Internet. Public cloud services may be free or offered on a pay-per-usage model.

2.2 Private Cloud

The private cloud infrastructure is operated for the exclusive use of an organization. The cloud may be managed by that organization or a third party. Private clouds may be either on- or off premises.

2.3 Hybrid Cloud

A hybrid cloud combines multiple clouds (private, community of public) where those clouds retain their incomparable identities, but are limit together as a unit. A hybrid cloud may pass standardized or ownership access to data and applications, as well as application portability.

III. LITERATURE REVIEW

Maninder et al.(2011) compared and evaluated Five search engines (Google, yahoo, bing, ask, AltaVista) on the basis of their search capabilities into two sections[5]. In the first section, features of five search engines are compared which are available to the user while searching the information. In second section, performance and capability is analyzed from the user's point of view. For this, they had taken a survey in which 263 participants participated and examined their interests in search engines. From this survey, they find out which search engine provides best utility and services to the user and most likely used by the people and they find out that users give highest rank to Google. Ya-Lan et al.(2007) proposed two major factors hygiene factor and motivation factor. Hygiene factors are those more fundamental requirements for a search engine and make users willing to use a search engine, and motivation factors are those more additional services of a search engine and make users willing to keep using the same search engine. The author had surveyed 758 people in Taiwan. The survey had three main components:

1) Demographic questions, the results showed that the age of 95% of the respondent's centers on the range from 18 to 30, and most of the participants are students

2) Experiences of using computer, Internet, and search engine, the results showed that more than 75% of the participants have experiences of using computer and Internet for more than five years. More than 95% of them use computer and surf on the Internet everyday for at least one hour.

3) Perceptions of search engines, test the hygiene-motivation hypothesis of search engine proposed in this research paper. Maninder et al evaluated five search engines but based on limited user review. Whereas Ya-Lan et al have used different factors for user liking and behavior, the results are dependent of various previous studies and the factors ought to take a unidirectional approach[6].

Rashid et al.(2009) devised an automatic web search evaluation system based on rough set based rank aggregation technique. Basically , different ranking results obtained from different techniques are combined. Two phases are used, ranking rules learning phase and rank aggregation phase.Author used 15 queries in rank learning phase. The output of this phase is a set of ranking rules[9].

George et al. (2007) suggested a model to evaluate search engines on the click through data of past users. The model used two variables i.e. A(attractively) and C(consideration) to determine the probability of choosing a snippet out of the list of relevant pages through which he successes to a distance d ; after considering upto distance d-1 portions. The conclusion of evaluation shows that the distance model represents the data better than popularity model. The complete evaluation illustrates that the positional biasing of relevancy can be resolved by click through data. Here it may seem counter-intuitive to use this model to measure performance. This toy model is unable to represent clearly the user behavior but it can be further improved to implement click through data methods [8].

IV. PROPOSED METHODOLOGY

The General idea behind this Paper is to create a system through which we can access the files which are present at different location on the remote systems. We mainly focus on the LAN based systems. The data flow diagram for the systems is shown:

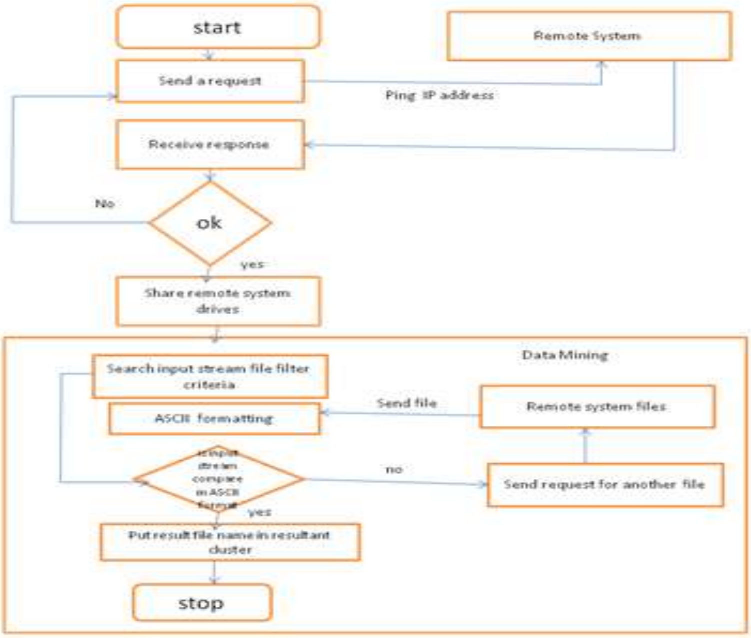
start
Remote System
Send a request
Ping IP address
Receive response
No
ok
yes
Share remote system drives
Data Mining
Search input stream file filter criteria
Send file
ASCII formatting
Remote system files
is input stream compare in ASCII format
no
Send request for another file
yes
Put result file name in resultant cluster
stop

Fig 2: The data flow diagram for the systems

The working can be done in following steps:

1. First we enter inside the systems using proper Username & Password.
2. After entering inside the system we first ping the IP address to the remote systems. This procedure is done with the help of sending request to the Remote Systems.
3. After sending request to the remote systems we are waiting for the response. If the response is valid we share the systems otherwise we retry by sending the request to the remote systems.
4. Once we share the systems it is easy to share the drives of the different systems.
5. After sharing all the drives we start to search the files.
6. For searching the files on the remote systems we are using the concept of Data mining.
7. First we have given the input string which is nothing but the file which we want to search.
8. Remote system sends the files in the ASCII format.
9. Once we get the files from the remote system in the ASCII format we check whether the given input string search contents are in ASCII format or not?
10. After getting the input search contents of file in ASCII format we put the collection of all the resultant files in the form of cluster.

V. CONCLUSION

It has been shown that an LAN Based Search engine collecting full-text and metadata from homogeneously structured information sources can be successfully implemented by integrating a system through which we search the files from the remote systems easily. Content classification and metadata extraction have been shown as valuable methods for enhancing search results. The System architecture provides us the strategies to enhance the retrieving the information from the remote systems. This system makes an attempt to propose a solution to retrieve higher occurrence of the keywords/concepts, within the files that are present on the different systems. Using this system we reduces the effort made by the user, that is, without requiring manual refinement, Where the relation-based metadata, provides a relevance score for a web page into annotated result set on user query, and the page annotation, and also decreases the time complexity.

REFERENCES

[1] Chuanping Hu, Zheng Xu, Yunhuai Liu, Lin Mei, Lan Chen, and Xiangfeng Luo,"Semantic Link Network based Model for Organizing Multimedia Big Data", IEEE Transactions on Emerging Topics in Computing(2014).

[2] Luis M.Vaquero,Luis Rodero-Merino, Juan Caceres, Maik Lindner," A Break in the Clouds: Toward a Cloud Definition", ACM SIGCOMM Computer Communication Review, 2009, 39(1): 50-55

[3] Abdur Chowdhury, Ian Soboroff, "Automatic Evaluation of World Wide Web Search Services",ACM,pp.421-422(2002)

[4] Himanshu Sharma, Bernard J. Jansen, "Automated Evaluation of Search Engine Performance via Implicit User Feedback" The Pennsylvania State University, ACM pp. 649-650 (2005).

[5] Maninder Kaur, Nitin Bhatia, Sawtantar Singh," Web Search Engines Evaluation Based on Features And End-User Experience", International Journal of Enterprise Computing and Business Systems,Vol. 1 issue 2(2011).

[6] Ya-Lan Chuang, Ling-Ling Wu,"User-Based Evaluations of Search Engines: Hygiene Factors and Motivation Factors, National Taiwan University, Proceedings of the 40th Hawaii International Conference on System Sciences, pp. 1-10(2007).

[7] Gordon & Pathak,"Finding information on the World Wide Web: the retrieval effectiveness of search engines". Information Processing and Management , pp. 141-180(1999).

[8] Georges Dupret ,Vanessa Murdock, Benjamin Piwowarski, "Web search evaluation using click throughdata and a user model (2007).

[9] Rashid Ali,M.M. Sufyan Beg,"Automated Performance Evaluation of Web Search System using rough set based rank aggregation" Proceedings of the first international conference on Intelligent Human Computer Interaction, Springer, pp.344-358(2009).