International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 7–July 2013

Identification of Effective Public Cloud on User Query

Manjeet Gupta[1],Sonia Sindhu[2]

[1]*Assistant Professor , Department of Computer Science & Engineering, Seth Jai Prakash Mukund Lal Institute of Engineering &Tech , Radaur, India .*

[2]*Mtech pursuing , Department of Computer Science & Engineering, Seth Jai Prakash Mukund Lal Institute of Engineering &Tech , Radaur, India .*

***ABSTRACT*- One of the major requirement over the web is about the selection of best service and service provider over the web. When we talk about cloud service the work is more specific and the parametric. As the cloud computing is becoming prevalent, a huge amount of information is centralized on the cloud so the searching of effective cloud data according to user requirement is necessary. A ranked search greatly improve the usability by returning the matching files in a ranked order. The ranking of returned data can be decided by several factors. Ranking of returned data based on keyword matching is already given by several authors. Here we will give ranking based on other factors like user recommendation ,response time and security as well. We describe algorithm for searching cloud data in which we can give different weight age to all these parameters. Here we are using public cloud for showing searching results.**

Keywords: Keyword search, relevancy, visitor count, crawler.

I.INTRODUCTION

The Cloud Services is the new trend of computing where readily available computing resources are exposed as a service. These computing resources are generally offered as pay-as-you-go plans and hence have become attractive to cost conscious customers .Cloud computing have three service models which are software as a service (Saas) , platform as a service (Paas) and infrastructure as a service (Iaas). Therefore cloud data searching scheme is performed for all these service models. Cloud can be deployed by using four deployment model , private cloud ,public cloud , community cloud and hybrid cloud. Here we are using public cloud for our searching scheme. Search engines use web crawlers to collect information about what is available on public web pages. Their primary purpose is to collect data so that when Internet surfers enter a search term on their site, they can quickly provide the surfer with relevant web sites. When a search engine's web crawler visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich Meta tags. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information.[1-3] The website is then included in the search engine's database and its page ranking process. The Cloud Service crawler collects information about the website and it's links.

- the website url
- the Cloud Service page title
- the meta tag information
- the Cloud Service page content
- the links on the page, and where they go to.

II.RELATED WORK

The problem of information retrieval has been discussed by several authors but they take only keyword matching parameter. In Year 2009, Georgia Koutrika presented a data cloud in which cloud search is performed on the basis of query summarization approach. The work presented by the author is a structural work in which the keyword extraction and the summarization is performed by the researcher and on the basis of this navigation and visualization of the data is suggested.[6]-[7].In Year 2012, Cengiz Orencik presented a rank based keyword search on the data cloud. In this work the document retrieval is performed on the cloud server based on the keyword analysis and the information search is performed relative to the defined information. The presented work is performed on the encrypted data that has improve the

security.[8]Several protocols have also used to maintain the security in ranked search in cloud computing. One of them is private information retrieval(PIR) ,provides useful cryptographic tools to hide the queried terms and the the data retrieved from the database while returning most relevant documents to the user[9]. With the growth of music collection ,music information retrieval (MIR) has been given in recent years. There are several ways to retrieve pieces of desired music.For example query by meta-information and query by tag. In content based MIR system ,user input a query of multiple tags with multiple level of preference by colorizing desired tags in a web based tag cloud interface to search music[10].Keyword search of PubCloud is also used in PubMed (database of biomedical literature). PubMed which is part of National Center for Biotechnology Information (NCBI), is a centralized database that indexes millions of biomedical publications. Responses to queries are presented by ranked lists that are similar to responses of most web search engines[15].

Another tag based summarization approach is suggested for the web search. The presented work is suggested on the public cloud. In which the integration of the web architecture and the database extraction is integrated.[11] The work includes the refinement of the user query based on the cloud tags. The words extracted from the query are been summarized and this summarized query is passed to the public cloud. The cloud interface enabled the extraction of new and required information.[12]-[13]-[14].

III.FRAMEWORK

The common steps involved in proposed work to search data is presented in the following subsections.

Step1. Cloud Service crawling

Step2. Cloud Service document parsing

Step3. Stop word removal

Step4. Similar score calculation

Step4. My search engine

Step6. Displaying search result

A . Cloud Service Crawling

First step in detection of duplicate and near duplicate Cloud Service pages is Cloud Service crawling .The analysis of the structure and informatics of the Cloud Service is facilitated by a data collection technique known as Cloud Service Crawling. The collection of as any beneficiary Cloud Service pages as possible along their interconnection links in a speedy yet proficient manner is the prime intent of crawling. Automatic traversal of Cloud Service sites, downloading documents and tracing links to other pages are some of the features of a Cloud Service crawler program.

Numerous search engines utilize Cloud Service crawlers for gathering Cloud Service pages of interest besides indexing them. Cloud Service crawling becomes a tedious process due to the subsequent features of the Cloud Service, the large volume and the huge rate of change due to voluminous number of pages being added or removed each day. Seed URLs are a set of URLs that a crawler begins working with. These URLs are queued. A URL is obtained in some order from the queue by a crawler. Then the crawler downloads the page.

This is followed by the extracting the URLs from the downloaded page and enquiring them. The process continues unless the crawler settles to stop. A crawling loop consists of obtaining a URL from the queue, own loading the corresponding file with the aid of HTTP, traversing the page for new URLs and including the unvisited URLs to the queue.[4-5]

B. Cloud Service Document Parsing

After Cloud Service crawling parsing of Cloud Service pages is done. Parsing splits a sequence of characters or values into smaller parts. It can be used for recognizing characters or values that occur in a specific order. In addition to providing a powerful, readable, and maintainable approach to regular expression pattern matching, parsing enables you to create your own custom languages for specific purposes. Information extracted from the crawled documents aid in determining the future path of a crawler. Parsing may either be as simple as hyperlink/URL extraction or complex ones such as analysis of HTML tags by cleaning the HTML content. It is inevitable for a parser that has been designed to traverse the entire Cloud Service to encounter numerous errors. The parser tends to obtain information from a Cloud Service page by not considering a few common words like a, an, the and more, HTML tags, Java Scripting and a range of other bad characters.

C. Stop Word Removal

Stop words are common words that carry less important meaning than keywords. Usually search engines remove stop words from a keyword phrase to return the most relevant result. I.e. stop words drive much less traffic than keywords. It is necessary and beneficial to remove the commonly utilized stop words such as "it", "can" ,"an" y "and", "by", "for", "from", "of', "the", "to", "with" and more either while

parsing a document to obtain information about the content or while scoring fresh URLs that the page recommends. This procedure is termed as stop listing. Stop listing aids in the reduction of size of the indexing file besides enhancing efficiency and value.

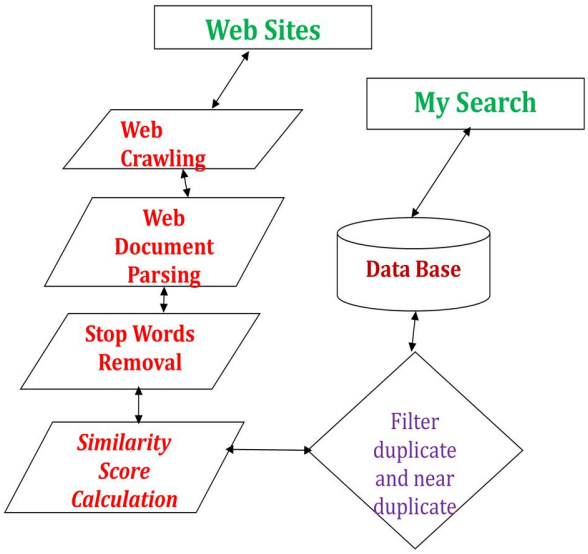

Figure 1 Represents the implementation of result

D. Similarity Score Calculation

If the prime keywords of the new Cloud Service page do not match with the prime keywords of the pages in the table, then the new Cloud Service page is added in to the repository. If all the keywords of both pages are same then the new page is considered as duplicate and thus is not included in the repository. If the prime keywords of new page are same with a page in the repository, then the similarity score between the two documents is calculated**.**

E. Step5: My Search

This provides a search engine like any popular search engine. But the result is fully depended for crawled Cloud Service URLs. This module demonstrates the removed duplicate and near duplicate page result.

F. Displaying Search Result

If the prime keywords of new page are same with a page in the repository, then the similarity score between the two documents is calculated. The keywords in the tables are considered individually for the similarity score calculation. If a keyword is present in both the tables, the formula used to calculate the similarity. The Cloud Service documents with similarity score greater than a predefined threshold are near duplicates of documents already present in repository. These near duplicates are not added in to the repository for further process such as search engine indexing.

G.Ranking Method

Term Weightage(c1)	User recommendation(c2)	Response time(c3)	Security vector(c4)

We will mark these four vectors on the scale of 10.

Example: For URL A

1) Term Weight-age: In case if term weight-age is 55% for A, then we will count it as 5.5

2) User Recommedation

a)Visitor Count: If visitor count is say 4 then, it will be considered as 0.4

b)Like/Dislike:

Like Count: 4

Dislike Count: 3

Then, Like-Dislike=1 on the scale of 10, we will mark it as 0.1

If (Like-Dislike is in negative), then we will use it with a negative sign.

If Like=3

Dislike=5

Then Like-Dislike=-2

Then we will mark it as -0.2

Page Rank: If page-rank is 60 then, it will be considered as 6

Now, using these terms we will calculate the rank of URL

w1*c1+w2*c2+w3*c3+w4*c4= Rank of URL A

Where,

w1= 0.3, w2=0.4, w3=0.27, w4= 0.03 (These are tested values).c1,c2,c3,c4 are the above defined parameter's values.

ALGORITHM

1.Define the list of available clouds on any public cloud server called Cloud(1),Cloud(2)…..Cloud(n)

2.Repeat For i=1 to n

Identify parameters for Cloud(i) called Availability(i),ResponseTime(i), Security(i).

3.Accept the UserQuery called Req under the specificationReqKeyword,ReqSecurity,ReqDeadLine ,

4.Activate the Middle layer to provide the best service selection

5.Accept the user query and filter it to retrieve the keywords under the following step:

a.)Remove the stoplist words from the query list.

b.)Rank the different keywords respective to category.

c.)Find the frequency of keywords.

d.)Keep the most occurring keywords and present as relevancy measure.

6.As the keywords retrieve perform query on each public cloud and perform the content and tag based match.

7.Find the list of M clouds that satisfy the relevancy criteria as well as identify the other cloud parameters like response time, security measure

8.Repeat For i=1 to M

[Perform the Content based similarity measure]

9.RelevencyVector = 0

Repeat For j=1 to Length(UserKeywords)

RelevencyVector=RelevencyVector+ KeywordOccurance(Cloud(i) ,Keyword(j))/TotalKeywords(Cloud(i),Keyword(j));

10.SecurityVector=0;

If(UserSecurityReq=Security(Cloud(i))

SecurityVector=1;

11.ResponseTimeVector=0

If(UserDeadline>ResponseTime(Cloud(i))

ResponseTimeVector=UserDeadline-ResponseTime(Cloud(i);

12.Rank(Clound(i)) = RelevencyVector*w1 + SecurityVector*w2 + ResponseTimeVector*w3+user recommendation*w4;

13.As user get Ranked list of clouds, selection can be performed for best cloud service provider respective to user interest.

IV.RESULTS

We can show the results by the following diagram :

Figure2 shows output for the user query for education

Here we have shown that if we increase the likes the ranking of url increase and if we increase the dislikes the rank decreases. The rank also depend on visitor count. All these factors shows user recommendation .The work has been implemented in java. In this work, the GoogleApps is used as the public cloud repository to perform the query analysis.

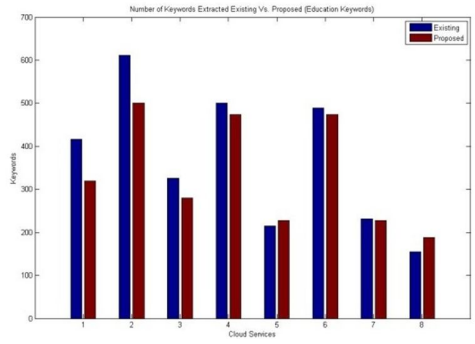

Figure3 Comparison of Keyword Search Education (Proposed Vs. Existing)

. The existing work represents the query performed on the cloud search without keyword extraction where as the proposed approach shows the keyword extraction after the keyword analysis. Here the outcome of the keyword analysis of education is shown. As we can see, the presented approach given more filtered relevancy so that the comparison can be performed easily.

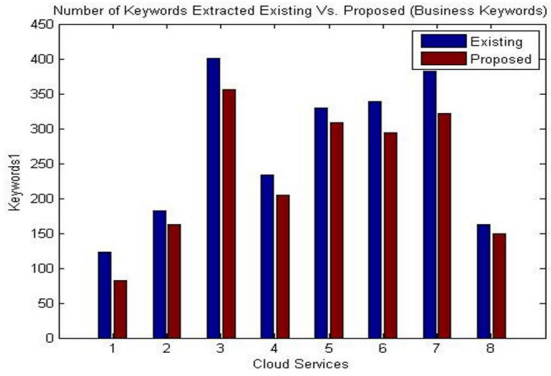

Figure4 Comparison of Keyword Search Business (Proposed Vs. Existing)

V.CONCLUSION

The presented work is about to perform the effective search in the cloud environment based on the user query relevancy factor. The relevancy of the query is here analyzed under three main factors called Keyword based Analysis, User Recommendation Analysis and the User Cloud service visit analysis. Based on these all factors a ranking criteria is decided and based on this ranking vectors the cloud services are ordered. The user can get the best cloud service as well as recommend other for the best service selection. In this work, the GoogleApps is used as the public cloud repository to perform the query analysis. The work is implemented in a web environment to perform the user query and to derive the ordered results from the query. In this present work, the work is performed on Google App engine for the public cloud repository but in future we can also work on other cloud servers or in the open environment. The efficiency in this work is the major issue, as the repository is larger and not having the service database, the search contents takes time. In future, the work can be improved in this direction.

REFERENCES

[1] http://en.wikipedia.org/wiki/Webcrawling
[2] http://www.encyclopedia.com/topic/Webcrawler.aspx.
[3] http://en.wikipedia/wiki/Distributed_web_crawling.
[4] http://www.wisegeek.org/what-is-a-web-crawler.htm.
[5] www.mendeley.com/catalog/web-crawling.
[6] Tritty Mamachan, Roshni .M. Thanka,"Servey on Keyword Searching in Cloud Storages .International Journal of Emerging Technology and Advanced Engineering(2012).
[7] Georgia Koutrika (2009)," CourseCloud: Summarizing and Refining Keyword Searches over Structured Data", EDBT 2009, March 24–26, 2009, Saint Petersburg, Russia. Pp. 1132-1135
[8] Dimitrios Skoutas (2011)," Tag Clouds Revisited", CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK. Pp 221-230.
[9] Cong Wang, Ning Cao, Jin Li(2011)," Secure Ranked Keyword Search over Encrypted Cloud Data".
[10] Ju-Chiang Wang (2011)," Colorizing Tags in Tag Cloud: A Novel Query-by-Tag Music Search System", MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.ACM p 293-302.
[11] Cengiz Orencik (2012) ," Efficient and Secure Ranked Multi-Keyword Search on Encrypted Cloud Data", PAIS 2012, March 30, 2012, Berlin, Germany. ACM, p 186-195.
[12] JinLi,Qian Wang(2010),"Fuzzy Keyword Search over Encrypted Data in Cloud Computing",Mini-Conference at IEEE INFOCOM2010.
[13] Remya Rajan(2012),"Efficient and Privacy Preserving Multi User Keyword Search for Cloud Storage Services" ,International Journal of Advanced Technology & Engineering Research (IJATER), Volume2,Issue4,July 2012.
[14] FuKuoTseng,SRongJaveChen(2012) "Toward Authenticated and Complete Query Results from Cloud Storages" IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communication.
[15] Byron Y-L. Kuo," Tag Clouds for Summarizing Web Search Results", WWW 2007, May 8–12, 2007, Banff, Alberta, Canada. ACM 978-1-59593-654-7/07/0005.