International Journal of Database Theory and Application
Vol.7, No.6 (2014), pp.223-232
http://dx.doi.org/10.14257/ijdta.2014.7.6.20

Use of Edit Distance Algorithm to Search a Keyword in Cloud Environment

Riya Mary, A.Sayali Nishikant, C. Jaya subalakshmi R and N. Ch. S. N. Iyengar

School of Computing Science and Engineering
VIT University, Vellore, India- 632014
riyamaryabraham@gmail.com,sayali31889@gmail.com,
suba_2026@yahoo.co.in,nchsniyr@vit.ac.in

Abstract

Cloud is a heterogeneous group of services and one of that is data storage. Generally the data stored in cloud is in very large amount, retrieval of the same requires precision and accuracy. Many methods are being proposed to retrieve data from cloud and strict match algorithm is one of them. In this paper we use of Levenshtein distance to find out the amount of similarity between two different strings. We compare the keyword given by user with our set of words, find out similarity and provide guesses to the user as accurate as possible.

Keywords: *Levenshtein distance, keyword search, similarity, cloud environment*

1. Introduction

Cloud Computing has become an integral part of technology, which is a type of computing that provides service over a network. It is based on the sharing of resources and the effectiveness in sharing. In cloud computing, without purchasing license a single server can be accessed by multiple users to retrieve and update their data for different applications. The term "moving to cloud" can be interpreted as moving away from traditional dedicated hardware systems to shared cloud infrastructure which works on the principle of pay as one uses it. Cloud computing allows to avoid upfront infrastructure costs, and helps on projects that give importance to their businesses rather than infrastructure. It helps IT industry to adapt with the changing business requirements and allows enterprises to get their applications up and running faster, with improved manageability and less maintenance. As cloud is based on "pay as you go" model, proper pricing measures need to be chosen to avoid unexpected charges.

The factors which lead to the evolution of cloud computing are virtualization in hardware, SOA and other computing methods such as grid and utility computing. The goal of cloud computing is to provide users a method or service to access the maximum from all of these technologies, the main benefit is that it can be achieved without the need for deep knowledge about or expertise with each one of them. The cloud aims to reduce costs, and help the users focus on their core business.

Virtualization is the fundamental technology for cloud computing. Virtualization software allows a physical computing device to be electronically separated into one or more "virtual" devices, each of which can be easily used and managed to perform computing tasks. This builds scalable systems based on multiple independent devices for computing.

Cloud Computing provides services based on the fundamental service models namely

- Infrastructure as a Service (IaaS)

ISSN: 2005-4270 IJDTA

- Platform as a Service (PaaS)
- Software as a Service (SaaS)

Deployment model in a cloud can be of different kinds such as

- Private – It is mainly for a single organization ,it can be operated either internally or with the help of the third party
- Public – It is a kind in which if the services are open for public use. It can be either free or based on pay and use model
- Hybrid – As the name suggests it is the combination of two or more clouds and provides the benefits of various models

Cloud Computing is one of the hottest topic of research and discussion. Due to the emergence of the cloud based or shared resource based model system, our day to day life data are stored in the cloud. It helps the data owners from the problems related to storage and maintenance. In cloud the amount of data stored is huge; the users require only the data of their specific need out of the whole. One of the most effective ways to retrieve only the required data is to perform a keyword based search.

In this work we propose a system which uses a strict matching algorithm which to search word in the cloud. It uses Levenshtein Distance method to find out the most similar words with the word entered by the user and provide convenience to the user.

Our work contains a dataset of medical information and search over any of the patient's information.

2. Literature Survey

Cloud Computing provides service over network by means of virtualization of hardware and the storage of huge amount of data. In order to access specific information from the data stored efficient search algorithms should be used. Several related works in this field has been carried out. This section explains the existing methods for searching a specific data from the cloud environment.

A method has been defined to the effective secure ranked keyword search over encrypted cloud data [1]. Order preserving encryption technique has been used. This method outperformed the available search techniques. In order to retrieve the required data keyword as well as concept based search has been used. Result was retrieved based on the broader conceptual entities. It has been used to overcome the drawbacks of the traditional search techniques on large amount of data.

The work has been carried out to effectively retrieve a user query from the whole data set [2]. A framework has been used for keyword searching with summaries which uses a ranking algorithm for ranked keyword search and their results. [3] Byron Y. L. Kuof defined a tag based summarization approach for searching in web which was carried out in public cloud. In his work the refinement of the user query based on the cloud tags have been presented. Summarized query has been used to extract the required information from the cloud

Ju-Chiang Wang proposed a content oriented tag based search [4]. For the analysis of query in the database music database was selected. In order to have an efficient content based search query analysis is done and categorized to different levels. The work has been evaluated for the effectiveness of the user query and the results obtained. Another work related to the searching in cloud proposed by Daniel E. Rose which depends on the information retrieval [5]. The work has been conducted on Amazon cloud service.

The work performed has been tested under different measures as well as in various environments including the enterprise level and the web level. It provided a more efficient way for content oriented search. Cengiz Orencik suggested a rank based keyword search on the cloud data [6]. The retrieval of the document is done based on the keyword search. The work has been performed on encrypted data and as a result required a secure method called private information retrieval. On this basis a secure protocol is suggested called Private Information Retrieval. Result obtained depends on the parametric ranking.

3. Analysis of Existing Techniques

Cloud provides a way to store and manage the large amount of data using the virtualization techniques and the shared resources aspects. Several techniques have been utilized in order to retrieve the required information from the cloud. Analysis on certain already existing techniques has been done in this section.

3.1 Ranked Keyword Search

In this approach as the first step, the user inputs the query for the extraction of information from the cloud and the system extracts the keywords from the input query. ie Query analysis is performed. This extracted keyword will be the input for the search architecture of cloud and based on the ranking algorithm it provides a ranked cloud list. In the ranking algorithm, it defines the list of available clouds; the system identifies parameters for the cloud. Then it accepts the query and performs the keyword extraction step which is followed by the query on each cloud based on the retrieved keyword. It gives a list of cloud which satisfies the criteria and based on the relevancy, response and security factors a ranked list of the cloud is returned. The method provides efficiency in query analysis and the results obtained but the major issue is the response time of the search.

3.2 Traditional Keyword Search

Traditional techniques allows to search the over data through keywords but it supports only Boolean search. Those techniques are not sufficient for the maximum utilization of the data which is present in the cloud in huge rate. The efficiency of this algorithm is very less. It has to scan the entire database for every keyword entered by the user. Hence this approach is not being used for keyword search.

4. Motivation

Cloud is an environment which stores the data which can be heterogeneous and provides different types of services to the users on “use and pay” basis. Cloud storage provides users with immediate access to a broad range of resources and applications hosted in the infrastructure of another organization via a web service interface. Cloud storage has several advantages over traditional data storage. For example, if you store your data on a cloud storage system, you'll be able to get to that data from any location that has Internet access. You wouldn't need to carry around a physical storage device or use the same computer to save and retrieve your information. With the right storage system, you could even allow other people to access the data, turning a personal project into a collaborative effort. So cloud storage is convenient and offers more flexibility. As cloud is meant to contain huge amount of data within it, whenever user wants to access a particular data from cloud, retrieval of that data needs an efficient searching method. We have made use of Levenshtein Distance method

to find out the most similar words with the word entered by the user and provide convenience to the user.

5. Architecture of System

As shown in the diagram above, the user uses the GUI we create in HTML, to input a word to be searched in database stored in cloud. The user is given suggestions for words present in our database using levenshtein's distance method. And based on the record user entered, all the other details of that record are displayed for the user on display screen.
It includes the following sections.

1. Creating database for the medical words
2. Creating ontology for the above database
3. GUI for user to search a keyword
4. Algorithm to find the word
5. Algorithm to find similar words

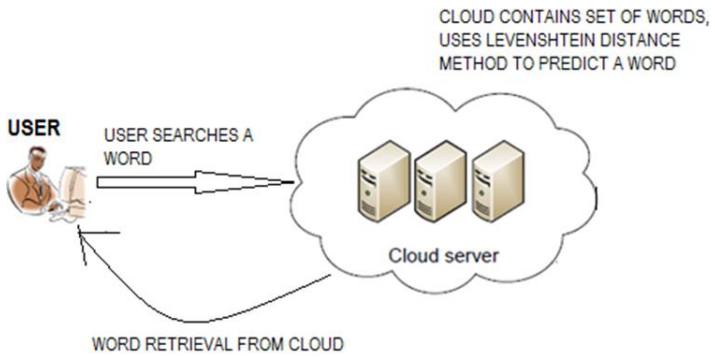

Figure 1. Architecture Diagram

5.1. Creating database for the medical words

We want our application to be used for medical purpose. Hence we create a database that contains unique ID for each patient, patient name, blood group, DoB, gender, diseases, measures taken.

5.2. Creating ontology for the above database

The ontology we create for the above database using the software protégé. We have created the light weight ontology for the same.

5.3. GUI for user to search a keyword

GUI is made with the help of HTML language and it contains a field to enter the keyword by user. A button to initiate the search and displays the result for all the patients data that is matching.

5.4. Algorithm to find the word

We make use of "strict match" algorithm to find the word from the huge database in least time possible. We make use of levenshtein distance algorithm here which is explained in following sections.

5.5 Algorithm to find similar words

1. User entered string is taken in variable s.
2. Each word in database is taken in variable t one at a time.
3. Length of both the strings is found out and maximum of the 2 is stored in variable max.
4. The edit distance between 2 strings s and t is found out and stored in variable d.
5. Similarity between s and t is found using the formula Sim= 1 - d/max
6. Similarity of user entered word for each word in database is found out and most similar word, *i.e.*, with highest similarity value is given as prediction to user.

Explanation with example

Let us consider for time being that the set of words stored in cloud are {January, February, March, April, May, June, July, August, September, October, November, December} Let the user wants to search some information for the month of August. Hence the user has to enter the keyword to be searched.

Let the word entered by user is "Augst". This word now will be compared with all the words present in our set. To find out most similar word with the word entered by user, we use levenshtein distance method. Here, we find the edit distance between each string and the string entered by user (say d) and similarity is found with the following formula

Sim= 1 - d/max of 2 strings

Hence for example, d for augst and January will be 5. Hence similarity will be 0.2857

Similarly for other strings,

Table 1. Similarity between 2 words

First String	Second String	Edit Distance	Similarity
January	Augst	5	0.2857
February	Augst	7	0
March	Augst	5	0.2857
April	Augst	5	0.2857
May	Augst	5	0.2857
June	Augst	4	0.4285
July	Augst	4	0.4285
August	Augst	2	0.7148
September	Augst	9	-0.28
October	Augst	7	0
November	Augst	8	-0.14
December	Augst	8	-0.14

As we can see in the table above, August has highest similarity with the string entered by the user, *i.e.*, augst and hence it will be provided to the user as prediction.

6. Mechanism Adopted and Tools Used

Cloud computing is the leading technology for delivery of reliable, secure, fault-tolerant, sustainable, and scalable computational services. It is not always possible to perform benchmarking experiments in repeatable, dependable, and scalable environments using real-world Cloud environments. A more viable alternative is the use of simulation tools. We make use of NetBeans to simulate our project. We have created the ontology using protégé and integrated the same with NetBeans as .rdf file. Also we have created a web services project in NetBeans which contains a GUI in html file and a backend code in java which searches the keyword entered by user in our database (*i.e.*, RDF file).

7. Results and Discussions

7.1 Lightweight ontology for the patients, doctors and nurse tables and their relations.

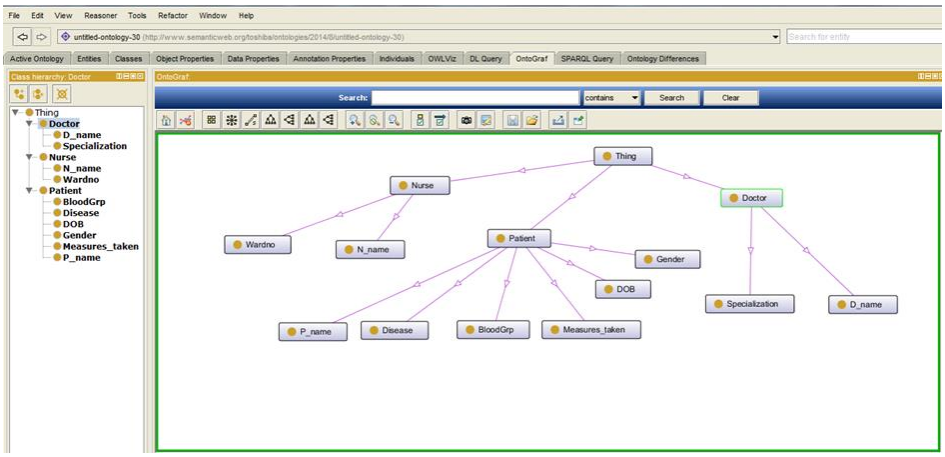

Figure 2. lightweight ontology created

7.2 The sample database we created for different patients with their details like blood group, date of birth, diseases they are suffering from and doctors assigned to them.

#	P_ID	P_NAME	DOB	BLOOD_GRP	DISEASE	MEASURES_TAKEN	ASSGND_DOC
1	4	Shalu Chandra	14-08-1992	B+	Jaundice	Ninco plus	Astha Malik
2	5	Ashish Patel	01-01-1995	A+	Jaundice	Ninco plus	Astha Malik
3	1	Anna Smith	03-04-1986	A+	Diabetis	Insulin	Chandrasekhar N
4	2	Robin Sharma	09-12-1976	O-	Heart Problem	Chrymo Medicine	Harsh Pathek
5	8	Apoorva Safai	09-04-1991	O+	Backpain	Moov,Excerise	Jacob Thomas
6	10	Gauri Salvi	03-08-1990	B-	Injuries	Bandage	Jacob Thomas
7	7	Madhura Brahme	08-04-1990	O-	Fever	Nimca plus	Shalini Devi
8	3	Meenakumari S	13-09-1990	AB+	Fever	Dolo650	Somali P
9	6	Gayatri Vijay	18-06-1989	A+	Cold	Paracetamol,Cetrezne	Sumith Yadhav
10	9	Lakshmi P	09-05-1992	A+	Fracture	Operation	Thamizharasi S

Figure 3. Database for patients

7.3 In HTML page, we take input from user like patients name and retrieve all the data.

The retrieved data for the value entered in above textbox is as shown below.

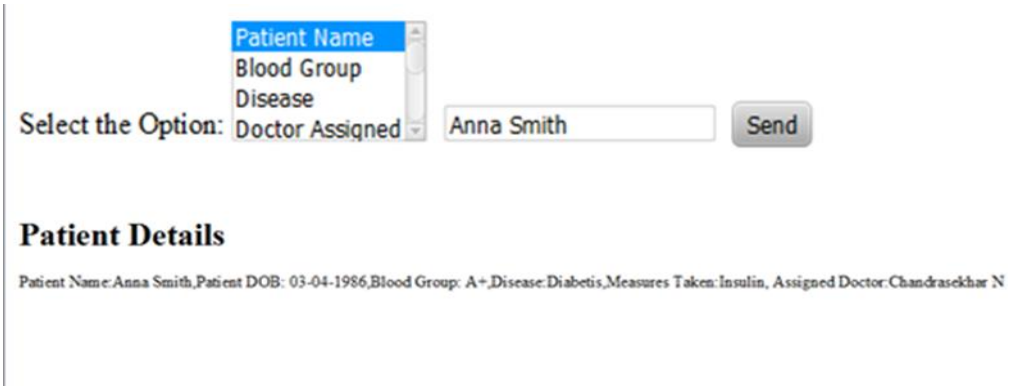

Figure 4. GUI design

7.4 We use Levenstein's Distance algorithm to check the similarity between 2 strings.

This can be used to provide predictions to user for the string he enters. Screenshot for the same is

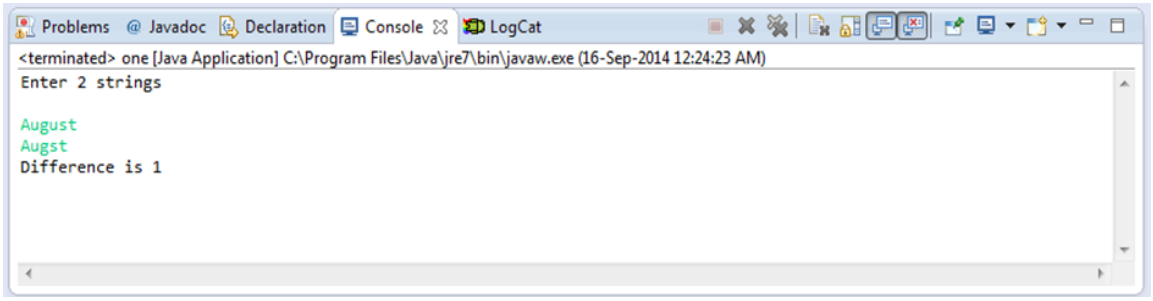

Figure 5. Algorithm output1

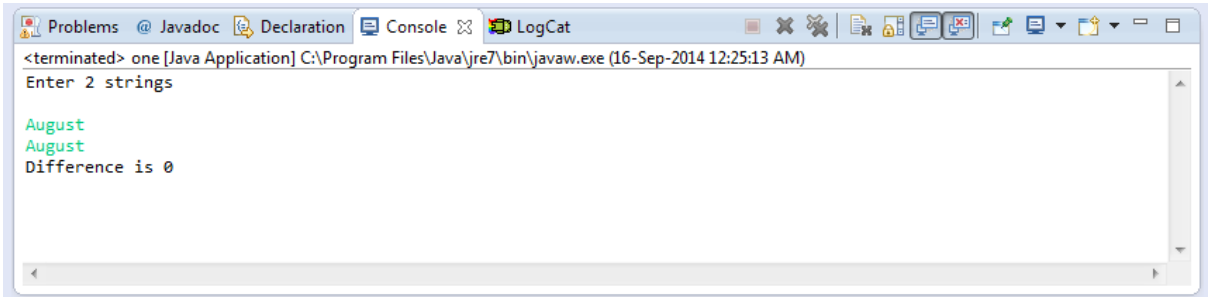

Figure 6. Algorithm output 2

7.5 Performance statistics

The graph below shows statistics for similarity predictions using levenshtein's algorithm and traditional search. As seen, we took 3 sets of words containing 23,53,21 words respectively and found out how many number of correct predictions using both algorithms for each set. We can take more than 3 sets for analysis too.

Table 2. Data for algorithm comparison

SET	Levenshtein's distance	Traditional Search
SET1(23)	22	18
SET2(53)	50	48
SET3(21)	21	18

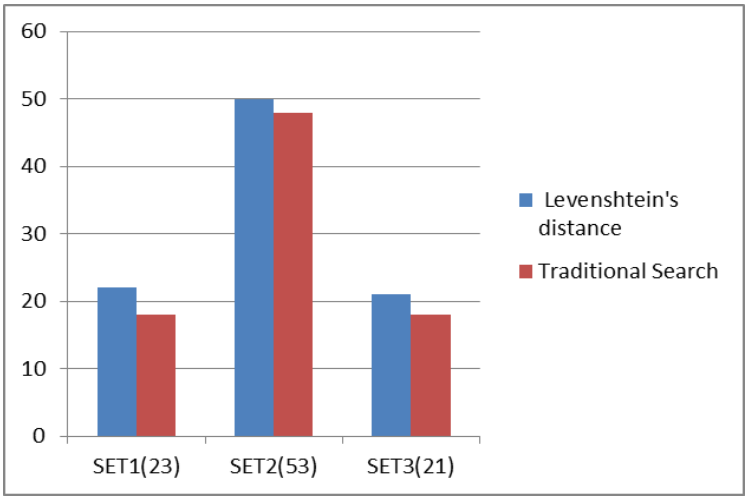

Figure 7. Graph showing Algorithm comparison

8. Conclusion

Cloud is a vast environment and today's trend to store data. We made use of Levenshtein's edit distance algorithm to provide predictions to user while he searches a word in cloud. Using this algorithm we could give more number of correct words compared to the traditional prediction mechanisms as shown in Figure 7.

Most of the information is obtained today by people by searching over internet. This method can be used in various search engines to give accurate predictions while user searches a particular string in cloud environment.

Our method gives predictions by word comparison; in future we plan to add more intelligence to it by considering the meaning of what user wants to search.

References

[1] M. R. Girme and G.M. Bhandari, "Efficient Secure Ranked keyword search Algorithms over outsource cloud data", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 2, no. 5,**(2013)**, pp-247-251.

[2] V. Makkar 2 and S. Dalal, "Ranked Keyword Search in Cloud Computing: An Innovative Approach", International Journal of Computational Engineering Research, vol. 3, no. 6, **(2013)**, pp 39-44.

[3] B. Y-L. Kuo, T. Hentrich, B. M. Good and M.D. Wilkinson, Proceedings of the 16th international conference on World Wide Web, **(2007)**; Banff, Alberta, Canada.

[4] J.-C. Wang, Y.-C. Shih, M.-S. Wu, H.-M. Wang and S.-K. Jeng, " Colorizing Tags in Tag Cloud: A Novel Query-by-Tag Music Search System", MM '11 Proceedings of the 19th ACM international conference on

Multimedia, **(2011)**; Arizona, USA.
[5] D. E. Rose, "Cloud Search and the Democratization of Information Retrieval", Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, **(2012)**; Oregon, USA.
[6] C. Orencik and E. Savas, "Efficient and Secure Ranked Multi-Keyword Search on Encrypted Cloud Data", Proceedings of the 2012 Joint EDBT/ICDT Workshops, **(2012)**; Berlin, Germany.

Authors

Riya Mary Abraham: She is currently pursuing post-graduation at VIT University Vellore,Tamil Nadu in Computer Science and Engineering stream. She has done bachelor of Technology in Computer Science and Engineering from College OfEngineering, Chengannur. Her major interest work area is Cloud Computing, Big data, Data warehousing and data mining.

Sayali Nishikant Chakradeo: She is currently pursuing post-graduation at VIT University Vellore, Tamil Nadu in Computer Science and Engineering stream. She has done Bachelor of Engineering in Computer Science and Engineering from Rajarshi Shahu College of Engg, Pune. Her major interest work area is Cloud Computing, Big data, Android development.

R. Jaya Subalakshmi is an Assistant Professor in the School of Computing Science and Engineering at VIT University, Vellore-632014, Tamil Nadu, India. She did M.S.(By Research) in VIT University. Her research area is Cryptography, Data Privacy and Agent based Distributed Computing.

Dr. N. Ch. S. N. Iyengar (b 1961) currently Senior Professor at the School of Computing Science and Engineering , VIT University, Vellore-632014, Tamil Nadu, India .He had 30 yrs of teaching experience. His research interests include Agent-Based Distributed secure Computing, Intelligent Computing , Network Security, Cloud Computing and Fluid Mechanics. He has authored several textbooks and had nearly 172 research publications in reputed peer reviewed International Journals. He delivered many keynote /invited lectures and served as PCM//TCM/reviewer for many International Conferences. He is Editor in Chief for International Journal of Software Engineering and Application (IJSEA) of AIRCC, Guest Editor for SI on Cloud Computing and Services of *Int'l J. of Communications, Network and System Sciences* and Editorial Board member for International Journals like **IJAST of SERSC, IJConvC of Inderscience** and many more.