Efficient Keyword based Text Mining of Documents in Cloud Environment

Geetika Saxena, Bharat Bhushan Agarwal
Computer Science Department
Iftm University
Moradabad

***Abstract:*-As we know we have to spend more time to search as well as to read the research and it takes more than two to three hours to read a single research paper, so it is necessary to use new search engine which provide best result because it is based on fastest reading algorithm. To give better verification of the research paper it will be very helpful.**

The traditional data mining assumes that information mined should always in the relational database form but in many cases information is available in the form of Natural Languages. The proposed work is thus based on the clustering and text mining in cloud environment. Clustering is widely studied data mining problem in text domain.

Keywords: Data Mining, Types of Databases, Cloud Computing, Cloud Models, Cloud Services, Cloud Components, Hierarchical Clustering, System Architecture

I. INTRODUCTION

Text mining is mining words from a baggage, also referred to as *text data mining. It* refers to the process of deriving relevant information from text or text data. The traditional data mining assumes that information minined should always in the relational database form but in many cases information is available in the form of Natural Languages. The proposed work is thus based on the clustering and text mining in cloud environment. Clustering is widely studied data mining problem in text domain. Saving the files to the cloud lets us access them from anywhere and makes it easy to share them with family and friends. Access your files from anywhere at any time, from any device.

The current working architecture uses a input of the keywords, after tokenization of these words, that work reading line by line and in this process is supported by a knowledge base that is fed into a first semi-automatically with the information collected from documents previously stored on the cloud server. Actually in this work cloud should be public in which information can be shared. In proposed work Text Mining is looking for the patterns in unstructured text. The problem finds the numerous applications in document management, document searching, document sharing, classification, visualization, collaborative filtering etc.

Text Mining is based on feature extraction, background knowledge; even patterns supported by small number of document may be significant. Information Retrieval (IR) systems identify the documents in a collection which match a user's query. As text mining involves applying very computationally-intensive algorithms to large document collections. For example, if we are interested in mining information only about clique graph theory, we might restrict our analysis to documents that contain the details about clique, or some form of the 'graph theory' or one of its synonyms.

1.1 Data Mining

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. Data mining is the discovery of knowledge from data. Thus, data mining should have been more appropriately named Knowledge mined from data, which is unfortunately somewhat long. Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data").

Fig. 1: Knowledge Mining

1. Removal of inconsistent data or to get normalized data.
2. Merge multiple data sources.
3. Selection of relevant data to the analysis task is retrieved from the database.
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance).
5. Data mining (an essential process where intelligent methods are applied in order to
   Extract data patterns)
6. Pattern evaluation
7. Knowledge representation techniques are used to present the mined knowledge to the user.The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

1.1 Types of database supported by data mining

- Relational Databases
- Data Warehouses
- Transactional Databases
- Object-Relational Databases
- Text Databases and Multimedia Databases

II. PROBLEM FORMULATION

In proposed work the simulator will show how to cluster documents from the different servers of the cloud on the basis of text mining. In this work Hierarchical Clusting Data mining algorithm is used. Effective Pattern Discovery for text mining and Tree Based mining for discovery patterns are the base papers they are working on text mining but there are chances of irrelevant search results. The proposed work is based on term based frequency approach which extract terms from the training set for describing relevant information there the relevant information refers to the best searched. The proposed approach is introduced to achieve excellent performance in the mining text files from the cloud. In this approach we will use cloud sim tool for cloud compating, many approaches were discovered to integrate data mining with information extraction.

III. CLOUD COMPUTING FUNDAMENTALS

Cloud Computing is internet-based computing service where shared resources like hardware, software and information are provided to the customers on demand, like the electricity grid.

Definition: Cloud Computing can be defined as a type of distributed and parallel system consisting of a collection of virtualized and inter connected computers [3]. Some of the emerging Cloud-based application services include data storage E-Banking, social networking, web hosting, content delivery, data processing etc.

Characteristics of Cloud Computing are as follows:

a) It is a type of client-server model such that clients are service requesters and servers are service providers.
b) There are heterogeneous types of servers available at service provider site to fulfill the varying demands of clients.

Cloud Computing is similar to utility Computing. The services are provided on account of the amount of resources used for the given time. Billing is done and the clients have to pay for the requested services.

- Location independent – Users can access systems using web browser regardless of their location and type of devices they are using.

Some of the examples of emerging Cloud computing infrastructures are Amazon EC2 [6], Amazon S3 [7], Microsoft Azure [8], Google App Engine [9], Aneka [10]. Having understood Cloud computing basics, its types need to be explored.

3.1 TYPES OF CLOUD MODELS

Cloud systems are divided into categories on the basis of the type of clients which will be taking its services. Different types of Cloud available are as follows:

a) Public Cloud Model: Public Cloud is a network of datacenters. Service providers of public Cloud sell services to anyone on the internet. The largest public Cloud provider is Amazon Web Services. Public clouds are owned and operated by companies that use them to offer rapid access to affordable computing resources to other organizations or individuals.

b) Private Cloud Model: Private Cloud is a proprietary network or a datacenter that supplies hosted services to limited number of people. An example is NASA's Nebula [11]. A private cloud is owned and operated by a single company that controls the way virtualized resources and automated services are customized and used by various lines of business and constituent groups.
   - On-Premise Private Cloud
   - Externally hosted Private Cloud

c) Community Cloud Model: Community Cloud came into existence when several organizations have similar requirements and seek to share infrastructure. It gives higher level of privacy as compared to public Cloud and its services are more expensive than public Cloud services. An example is Google's Gov Cloud [12].

d) Hybrid Cloud: Hybrid Clouds combine both public and private cloud models.

3.2 TYPES OF CLOUD SERVICES

3.2.1 INFRASTRUCTURE AS A SERVICE (IaaS) : Infrastructure as a Service (IaaS for short) is the backbone of cloud computing. Infrastructure as a Service is the first layer and using this service model, user can manage his applications, data, middleware operating system, and runtime.

3.2.2 PLATFORM AS A SERVICE (PaaS): This cloud service model could be considered the second layer . This is a variant of SaaS.

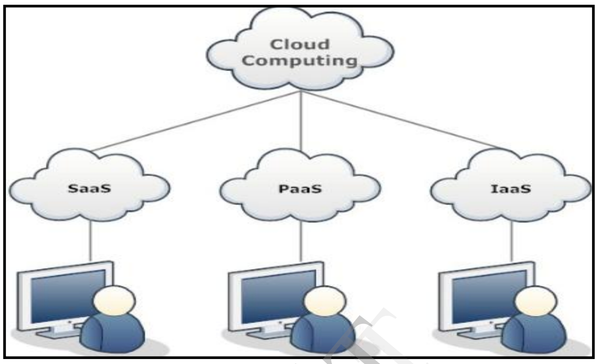

Fig. 2: Type of clouds Services

3.2.3 Web-Based Cloud Services: These services provide web services functionality not developed applications.

3.2.4 Software As A Service (Saas): This is the final layer of the cloud services model. These services provide support to run programs in the cloud.

3.2.5 Managed Services: This is perhaps the oldest iteration of cloud solutions.

3.3 Cloud Components

A cloud computing solution is made up of several elements clients, the datacenter and distributed servers. As shown in Figure, these components make up the three parts of a cloud computing solution.

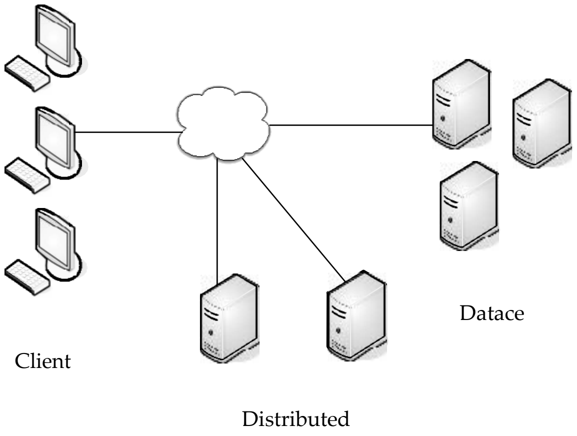

Fig. 3: Clouds Components

3.3.1 Clients

Clients are, in a cloud computing architecture, the everyday local area network (LAN). They are, typically, the computers on your desk. But they might also be laptops, tablet computers, mobile phones, or PDAs etc.

3.3.2 Datacenter

The datacenter is the collection of servers where the application to which you subscribe is housed. It could be a large room in the basement of your building or a room full of servers on the other side of the world that you access via the Internet.

3.3.3 Distributed Servers
3.3.4 Infrastructure

IV. METHODOLOGY USED

4.1 HIERARCHICAL CLUSTERING

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering (defined by S.C. Johnson in 1967) is this:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

4.2 Hierarchical Clustering Algorithm In Proposed System

In hierarchical clustering, there are two types of clustering, Agglomerative and Divisive as discussed in Chapter1. In proposed system we are implementing agglomerative approach. It starts with the points as individuals clusters. At each step it merge with the closest pair of clusters. Until only one cluster left (or k cluster left).

Agglomerative hierarchical clustering algorithm:

1. Initially each object forms it own cluster. In proposed system, objects are words in the documents.

2. Compute all the pair wise distances between the initial clusters(words) Repeat

3. Search the closest documents (A, B) in the set of the current clusters into a new cluster C=AUB. Here C is the folder for clustering A and B documents.

Copy A and B from the set of current clusters into C folder until only a single cluster remains.

Until

This single cluster or folder documents is saved in the client system.

According to agglomerative hierarchical clustering it starts with cluster of individual words by tokenizing the sentences using java code. A proximity matrix is also considered.

4.3 System architecture

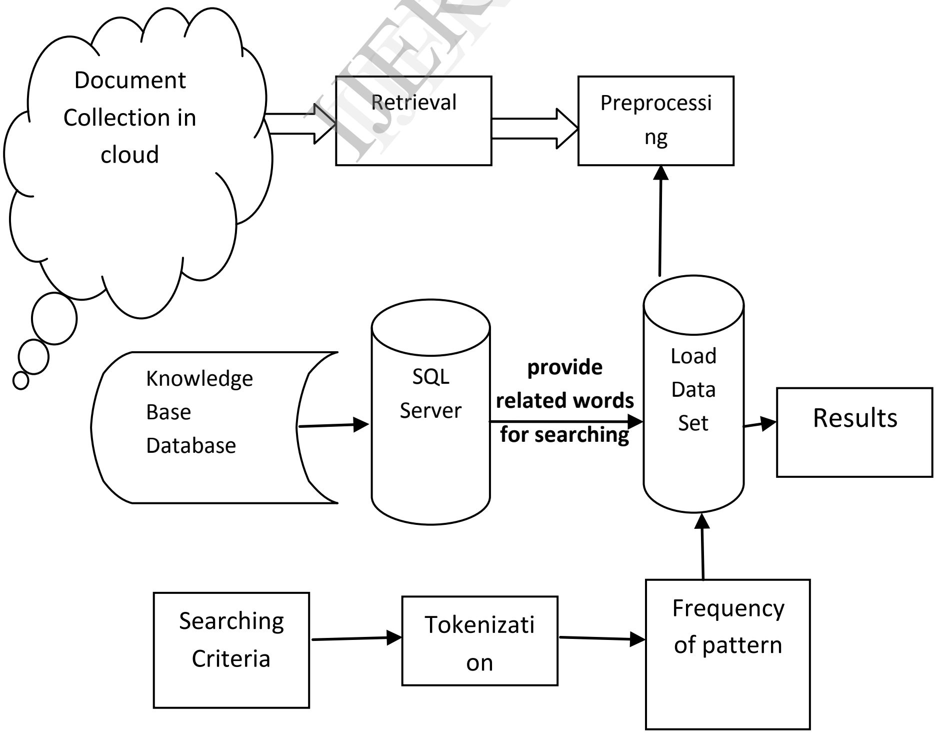

Fig.4: System architecture

V. CONCLUSION

The world has motivated to search for hidden knowledge in text collections because of the popularity of the Internet and the huge amount of documents are available on the internet. As the Internet provides various sources of useful information but it is difficult to access and extract their content. Information extraction (IE) software identifies and removes relevant information from texts, fetching information from different sources, and shows the combined view. Information extraction can be of two types: natural language processing and wrapper induction. The principal advantages of simulation are:

- Flexibility of defining configurations
- Ease of use and customization
- Cost benefits: It can be expensive for any application on the cloud to first designing, developing, testing, and then redesigning, rebuilding, and retesting. Simulations keeps out the building and rebuilding phase by using the model already created in the design phase.

Future Scope

In the recent years with the advancement of web and social network technology have lead to a tremendous interest in the classification of text document containing links or other meta- information tries to optimize accuracy, the efficiency and searching from wide area. It will provide intelligent text analysis.

REFERENCES

1. C.C. Aggarwal, Y.Zhao, P.S.Yu. On Text Clustering With Side Information, ICDE Conference,2012
2. C.C.Aggarwal, P.S. Yu. On Effective Conceptual Indexing and Similarity Search in Text, ICDM Conference, 2001.
3. C.C. Aggarwal, P.S. Yu. A Framework for clustering Massive Text and Categorial Data Streams, SIAM Conference on Data Mining, 2006.
4. C.C. Aggarwal, S.C. Gates, P.S. Yu. On Using Partial Supervision for Text Categorization, IEEE Transaction on knowledge and Data Enigineering, 16(2) 245-255, 2004.
5. C.C. Aggarwal, C. Projected, J. Wolf, P.S. Yu, J.—S. Park. Fast Algorithms for Projected Clustering, ACM SIGMOD Conference, 1999.
6. C.C. Aggarwal, P.S. Yu. Finding Generalized Projected Clusters in High Dimensional Spaces, ACM SIGMOD Conference, 2000.
7. R. Agarwal, J. Gehrke, P. Raghavan. D. Gunopulos. Automatic Subspace Clustering of High Dimensional Data for Mining Applications, ACM SIGMOD Conference 1999.
8. R. Agarwal, R. Srikant. Fast Algoritm for Mining Association Rules in Large Database, VLDB Conference, 1998.
9. J. Allan, R. Papka, V. Laverenko. Online new event detection and tracking. ACM SIGIR Conference, 2004.
10. P. Andritsos, P. Tsapars, R. Miller, K. Sevcik. LIMBO: Scalable Clustering of Categorical Data. EDBT Conference, 2004.
11. P. Anick, S. Vaithyanathan. Exploiting Clustering and Phrases for Context-Based Information Retrieval. ACM SIGIR Conference, 2004.
12. R. Angelova, S. Siersdorfer. A neighborhood-based approach for clustering of linked document collections. CIKM Conference, 2006.
13. R. A. Baeza-Yates, B.A. Riberio-Neto, modern Information Retrieval-the concepts and technology behind search, Second edition, Pearson Education Ltd., Harlow, England, 2011.
14. S. Basu, A. Banerjee, R.J. Mooney. A probabilistic framework for semi-supervised clustering ACm KDD Conference, 2004.
15. S. Basu, A. Banerjee, R.J. Mooney. Semi-supervised clustering, by Seeding. ICML Conference, 2002.
16. F. Beil, M. Ester, X.Xu. Frequent term-based text clustering, ACM KDD Conference, 2002.
17. L. Baker, A. McCallum. Distributional Clustering of words for text Classification , ACM SIGIR Conference, 1998.
18. R. Bekkerman, R. El-Yaniv, Y. Winter, N.Tishby. On Feature Distributional Clustering for Text Categorization. ACM SIGIR Conference, 2001.
19. D. Blei, J. Lafferty. Dynamic topic models. ICML Conference, 2006.
20. D. Blei, A. Ng, M. Jordan. Latent Diriclet allocation, journal of machine Learning Research, 3:pp. 993-1022, 2003.
21. P.F. Brown, P.V. deSouza, R.L. Mercer, V.J. Della Pietra, and J/C. Lai. Class-based n-gram models of natural language, Computational Linguistics, 18, 4 (December 1992), 467-479.
22. K. Chakrabatri, S. Mehrotra. Local Dimension reduction: A new Approach to Indexing High Dimensioanl Spaces, VLDB Conference, 2000.