ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 4, April 2016

Chemically Intelligent Text Mining System Using OpenStack

Kavya Kartha[1],Aishwarya Bhalerao[2],Vaibhavi Modake[3],Shraddha Waghmare[4],
Prof. Sudarshan S. Deshmukh[5] ,PCCOE
Computer Department, Savitribai Phule Pune University
India

***Abstract*-The cloud computing delivers hardware/software to the end user at a low cost with an easy to use interface via the internet. OpenStack is a free and open-source cloud-computing software platform. The aim of this paper is to show how a chemically intelligent text mining system implemented on OpenStack platform will help doctors for discovering chemical related relevant terms related to diseases. Basically text mining is done on medical database which refers generally to the process of extracting useful information from unstructured text. Text mining and knowledge extraction are ways to help researchers to handle large information overload . In the proposed system data is being retrieved and mined from a web source, PubMed which consist of large amount of medical information. This will analyse the chemicals related to a particular disease.**

***Keywords*-Cloud computing, OpenStack, Text Mining, PubMed, IaaS, Genia Corpus.**

I. Introduction

The increasing amount of medical data in medical databases requires certain technique that can analyze the data and extract knowledge from these databases. Cloud computing offers a variety of different service models in which the resources can be deployed :Infrastructure as a Service (IaaS), Software as a Service (Saas) and Platform as a Service(PaaS). IaaS (Infrastructure as a Service) is the low-level in the cloud, aims at providing compute and storage resources and virtual machines to users. Users of IaaS can completely control and configure their infrastructure resources as what they want.

OpenStack not only has following core projects: compute (Nova), storage (Swift), image (Glance), it also add two new projects: Dashboard (Horizon) and Identity (Keystone). All these projects indicate that OpenStack is striving to integrate more services into it so that provide a massively scalable and rich services cloud[2][9].

In this paper the proposed system will do text mining on medical database from unstructured text. The data is being retrieved and mined from a web source PubMed, which consist of large amount of medical information related to different diseases. This will analyze the chemicals related to a particular disease on OpenStack.

II. Literature survey

OpenStack is the collection of open source software projects that cloud providers can use to setup and run their cloud compute and storage infrastructure. The project aims to build an open-source community with researchers, developers and enterprises, they share a common goal to create a cloud that is simple to deploy, massively scalable and full of rich features[7].

Using cloud computing there is reduction in costs associated with managing hardware and software resources by allowing them to be multipurpose and easy for the end user to self manage. Openstack is full featured IaaS platform[7].This service model allows the resources to be deployed such as processing, storage and network resources. Using this hardware the end user can run operating systems or applications. The end user does not manage the underlying fixed infrastructure used to provide the cloud services but instead can control how the resources are deployed.

Components of OpenStack[5]:

1. Compute (Nova): Nova is a cloud computing fabric controller used to deploy and manage large number of virtual machines and other instances to handle computing tasks.
2. Object Storage (Swift): Swift is a scalable useful storage system for objects and files. Files are written to the large number of disk drives spread throughout servers in the data center. OpenStack software is responsible for ensuring data replication and integrity across the cluster.
3. Block Storage (Cinder): OpenStack Block Storage (Cinder) is a block storage component,more analogous to the traditional notion of a computer which can access specific locations on a disk drive and also it provides persistent block-level storage devices for using with OpenStack compute instances. In OpenStack, the block storage system is used to manage the creation, attaching, detaching of the block devices to servers.
4. Networking (Neutron): OpenStack Networking (Neutron) is used to provide the networking capability for OpenStack and it is a system which can

manage networks and IP addresses easily, quickly and efficiently.

5. Dashboard (Horizon): OpenStack Dashboard (Horizon) is the dashboard behind OpenStack which provides a graphical interface to access, provision and automate cloud-based resources for administrators and users .
6. Identity Service (Keystone): OpenStack Identity (Keystone) provides the identity services for OpenStack or it is a central directory of users mapped to the OpenStack services. It provides multiple means of access, and acts as a common authentication system .
7. Image Service (Glance): OpenStack Image Service (Glance) provide image services to the OpenStack, discovery, registration and also delivery services for disk and server images, it also allows these images to be used as templates when deploying new virtual machine instances.
8. Telemetry (Ceilometer): OpenStack Telemetry Service (Ceilometer) provides telemetry services, which will allow the cloud to provide billing services to individual users of cloud, it keeps a verifiable count of usage of each user’s system of the various components of an OpenStack cloud.
9. Orchestration (Heat): OpenStack Orchestration (Heat) is a service which facilitates developers to store the requirements of a cloud application in a file which defines the resources necessary for that application.

In the bioinformatics domain, biomedical research literature[10] has been a target for text mining. The first textbook on biomedical text mining with a strong genomics focus appeared in 2005 [3], where it has reported that industry has suggested that 90% of drug targets are derived from the literature.

The Text Mining Process is carried out in following stages[6]:

Stage-I: Pre-processing Text:

As compared to mining natural languages documents it is easy to mine from a pre-processed text .So before applying any text mining technique it is important to do the pre-processing.

Stage II- Text Mining Technique is applied:

This is an important stage in which an algorithm is selected and applied on text in order to process the text. Algorithm such as clustering, classification,information extractions or visualization can be done.

Stage III - Analysis of Text:

In this stage the outputs are analysed for discovering the knowledge.

Finding disease relationships requires laborious examination of hundreds of possible candidate heterogeneous factors.A data mining engine, namely MeSH Terms Associator (MTA), that has been employed in a distributed architecture to refine a generic PubMed query by means of discovery of concept relations in the form of association rules. Results show that MTA mines interesting rules.[1]

III. Proposed Work

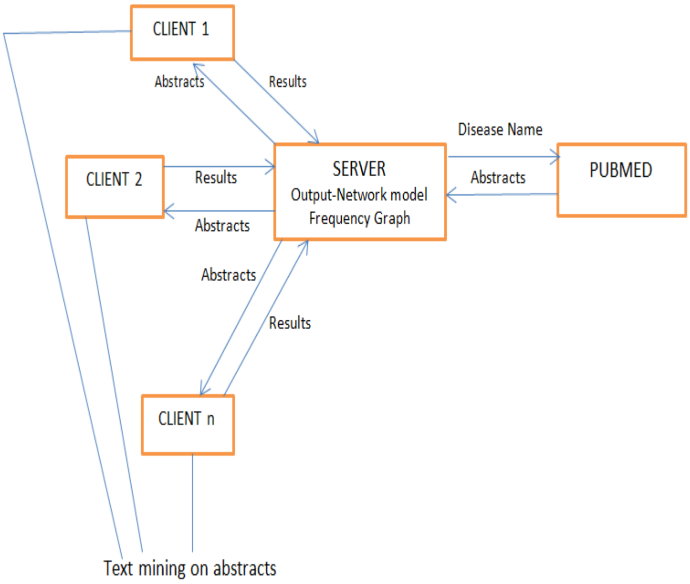

Fig 1: Proposed System Architecture

The proposed system in Fig 1. works as follows:

On the server side, we are basically going to divide the workload on the server i.e instead of performing the text mining on all the abstracts on the server side, we will divide the number of abstracts amongst the clients.

On the client side, it will receive the abstracts send by the server, text mining will be performed on the downloaded documents. The goal of text mining in this area is to allow biomedical researchers to extract knowledge from the biomedical literature to facilitate new discovery in a more effective manner[4].

As database can store less amount of information,this problem has been solved through Text Mining.Using the technique such as information extraction,the names of different entities and their relationship can easily be found from the corpus of documents set.

The result from the clients is sent back to server and the server performs the futhur analysis and builds a network model.

Steps for mining the required terms from the documents downloaded:

Step 1: Calculate the frequency count of each term within the abstract.

Step 2: Compare them with the medical dictionary terms which can be taken from Genia Corpus.

Step 3: Pick out the matching terms. Eliminate the false terms.

Step 4: Classify them into classes as per the requirements.

Step 5: Calculate the number of terms in each class

Step 6: Identify the class with the highest relevant terms that is more related to disease.

IV. Result

The frequency analysis graph is displayed as output on the screen. The graph of frequency verses the chemicals will be displayed. The class with the highest number of relevant terms will be related to that disease. Accordingly a network model depicting the relationship of the disease and those classes will be displayed on screen as well as shown in Fig 2.

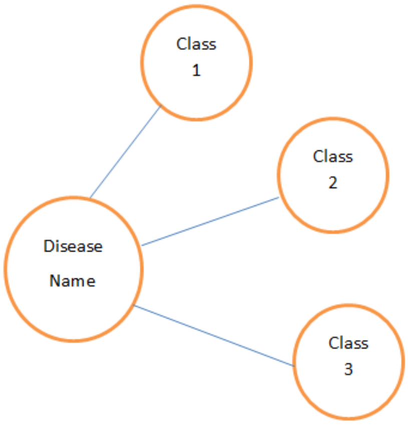

Fig 2: Network Model

V. Conclusion

The cloud computing paradigm has a great potential to provide a flexible platform for delivering computational resources. Openstack compute is designed for provisioning of virtual machines providing scalable cloud computing platform.Using OpenStack, we have implemented Infrastructure as a service cloud on Chemically intelligent text mining. Taking Disease as an input, we can analyse the relevant chemicals related to it in less provisional time due to virtual and distributed system. This system is truly useful for doctors and researchers for finding chemicals in less time which was previously done manually. We anticipate that this approach of deploying Cloud Computing Services to find chemicals helps the doctors and researchers to find anti-chemicals to cure the diseases.

Acknowledgment

We express our sincere thanks to our project guides Prof.Sudarshan S. Deshmukh and Dr. M. Karthikeyan for their encouragement and support throughout our project, especially for the useful suggestions given during the course of project and having laid down the foundation for the success of this work. We would also like to thank our for project coordinator Prof. Sonali Tidke and also HOD of computer department Prof. Dr. K. Rajeswari for their assistance, genuine support and guidance from early stages of the project.

References

[1] Margherita Berardi, Michele Lapi ,Pietro Leo ,Donato Malerba1 ,Caterina Marinelli, Gaetano Scioscia, "A data mining approach to PubMed query refinement",Proceedings of the 15th International Workshop on Database and Expert Systems Applications (DEXA'04),IEEE.

[2] Rohit Kamboj , Anoopa Arya, "Openstack: Open Source Cloud Computing IaaS Platform",International Journal of Advanced Research in Computer Science and Software Engineering,Volume 4, Issue 5, May 2014.

[3] Shaidah Jusoh and Hejab M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining",College of Computer Science & Information Systems, Najran University.

[4] W. Hersh, "Evaluation of biomedical text-mining systems: Lessons learned from information retreival", Briefings in Bioinformatics, vol. 6, no. 4, pp. 344–356, 2005.

[5] Rakesh Kumar,Neha Gupta,Shilpi Charu,Kanishk Jain,Sunil Kumar Jangir, "Open Source Solution for Cloud Computing Platform Using OpenStack",IJCSMC, Vol. 3, Issue. 5, May 2014, pg.89 – 98

[6] Divya Nasa, "Text Mining Techniques- A Survey",USICT , GGSIPU.

[7] Xiaolong Wen,Genqiang Gu,Qingchun Li,Yun Gao,Xuejie Zhang,"Comparison of Open-Source Cloud Management Platforms: OpenStack and OpenNebula",9th International Conference on Fuzzy Systems and Knowledge Discover,2012.

[8] Sachintha Pitigala, Cen Li, Suk Seo, "A Comparative Study of Text Classification Approaches for Personalized Retrieval in PubMed",IEEE International Conference on Bioinformatics and Biomedicine Workshops,2011.

[9] OpenStack Home Page. http://www.openstack.org/.

[10] Rodriguez-Esteban,R. (2009) "Biomedical text mining and its applications",PLoS Comput. Biol., 5, e1000597.