International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Data Mining and Assessment of Social Networking Site on Cloud

Sanchay Shrivastava
Computer Science
S.R.C.E.M., Banmore, (M.P.)
sanch4455@gmail.com

Aparajit Shrivastava
Dept of Computer Science
S.R.C.E.M., Banmore, (M.P.)
aparajit2k5@gmail.com

***Abstract*—The computing paradigm that comes with cloud computing has incurred great concerns on the security of data, especially the integrity and confidentiality of data, as cloud service providers may have complete control on the computing infrastructure that underpins the services. In this paper we want to generalize the formulation of data mining techniques with cloud computing environment. In data mining we want to find useful patterns with different methodology. The main issue with data mining techniques is that the space required for the item set and there operations are very huge. If we combine data mining techniques with cloud computing environment, then we can rent the space from the cloud providers on demand. This solution can solve the problem of huge space and we can apply data mining techniques without taking any consideration of space. Then we can perform data mining technique to find frequent patterns and relevant associations. After the completion of data mining technique we can deduce the cost in cloud and non-cloud environment.The role of data analytics increases in several application domains to cope with the large amount of captured data. Cloud computing has become one of the key considerations both in academia and industry. Cheap, seemingly unlimited computing resources that can be allocated almost instantaneously and pay-as-you-go pricing schemes are some of the reasons for the success of Cloud computing.**

***Keywords*—Cloud Computing, Data Mining, Social Network Analysis**

1. INTRODUCTION

Cloud are the virtualized resources supposed as large pools, which can be accessed easily. The resources of cloud are dynamically reconfigurable to gain optimum resource utilization. It is based on pay per use model, in which the client is allowed to use services of resources over cloud under the defined service level agreement (SLAs) by the service provider. In current scenario, there are number of individuals and organizations are taking advantage of powerful mass computing and storages centers, which are highly stable cloud architectures serviced by large companies. From a client perspective, cloud computing provides a means of acquiring computing services without requiring understanding of the underlying technology. From an organizational perspective, cloud computing services for consumer and business needs in a simplified way, providing unbounded scale and differentiate quality of service to foster rapid innovation and decision making. It is a service acquisition and delivery model for IT resources and control the costs of delivering IT resources to the organization. The benefits of cloud computing in providing access to digital content in developing countries. The utilization of cloud computing can reduce costs related with the IT infrastructure necessary for standard internet network services. This makes cloud computing very appealing to businesses and governments that lack the financial means to cover internet IT support and recurring hardware expenses. For example, in the education sector, many universities in developing countries use Software as a Service (SaaS) applications where cloud servers host virtual interfaces of educational content. SaaS applications are being offered by industry leaders such as Google (Google Apps for Education) and Microsoft (Outlook) to many Sub-Saharan and South African universities. With their cloud services they facilitate on-line courses, student emails, and payroll systems and improve the cross campus collaboration between various university departments. Because of the highly flexible nature of cloud services, universities and educational institutes are able to use some of them on demand, thus helping save funds which can be allocated for other purposes.

Another advantage of cloud computing related with SaaS educational applications is that it creates a learning environment that surpasses restrictions posed by time and location that some students face. Indeed, SaaS applications create an ‘any-time / all-year-round’ learning environment where students can log in to online blackboards, download course materials, take exams and tests or complete assignments more flexibly. It has also been argued that SaaS applications can actually increase the quality of curriculum and help students learn faster as they facilitate student interaction and make real time collaboration easy by bypassing geographical limitations set by location. K.P.N Jayasena believes that: ‘’Educators point to cloud computing as a solution for instructors’ obstacles in preparation and development of courses and strengthening curriculum. Material can be taught and absorbed much more quickly and easily through internet access, using cloud computing in the classroom ’’ In addition to that, cloud computing can increase interoperability between different institutions that use the same cloud provider therefore promoting joint research programs and exchange of important scientific information and databases between researchers.

2. CLOUD COMPUTING ENVIRONMENT

Microsoft, suggests that cloud computing is a metaphor for the internet. It is quite common these days to draw network diagrams that depict the Internet as a cloud hence the use of the word in this instance.

In a typical cloud computing scenario organizations run their applications from a data Centre provided by a third-party – the cloud provider [2]. It is cloud provider’s responsibility to providing the infrastructure, servers, storage and networking necessary to ensure the availability and scalability of the applications. This is what most people mean when they refer to cloud computing i.e. a public cloud.

Private clouds are very popular among companies or organizations. A private cloud is a proprietary computing architecture, owned or leased by a single organization, which provides hosted services behind a firewall to "customers" within the organization. Some commentators regard the term "private cloud" as an oxymoron. They say that the word "cloud" implies an infrastructure running over the Internet, not one hidden behind a corporate firewall. There is, however, a larger body of opinion suggesting that private clouds will be the route chosen by many large enterprises and that there will be substantial investment in this area. Already comapnies are lining up to release products that will enable enterprises to more easily offer internal cloud services.

There is a huge amount of hype cloud computing but despite this more and more C-level executives and IT decision makers consent that it is a real technology option. It has moved from innovative technology to a commercially viable alternative to running applications in-house. Vendor organizations such as Amazon, Google, Microsoft and Salesforce.com have invested many millions in setting up cloud computing platforms that they can offer out to 3rd parties. They clearly see a big future for cloud computing. Of course, no technology comes without a set of advantages and disadvantages so we've tried to sort to wheat from the chaff when it comes to the reality of cloud computing. In particular, one always has to be cautious in believing the claims of any specific vendor.

There are several reasons to approve cloud computing like Cost, Scalability, Business Agility, and Disaster Recovery. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [2].

This cloud model promotes availability and is composed of:

1. Five essential characteristics:
   - On-demand self-service;
   - Broad network access;
   - Resource pooling;
   - Rapid elasticity;
   - Measured Service;
2. Three service models:
   - Cloud Software as a Service (SaaS)
   - Cloud Platform as a Service (Platform as a Service)
   - Cloud Infrastructure as a Service (IaaS) and,
3. Four deployment models:
   - Private cloud;
   - Community cloud;
   - Public cloud;
   - Hybrid cloud.
4. Key enabling technologies include:
   - Fast wide-area networks;
   - Powerful, inexpensive server computers; and
   - High-performance virtualization for commodity hardware.

In general, a public (external) cloud is an setting that exists outside a company's firewall. It can be a service offered by a third-party vendor. It could also be referred to as a mutual or multi-tenanted, virtualized infrastructure managed by means of a self-service portal.

A private (internal) cloud reproduces the delivery models of a public cloud and does so behind a firewall for the exclusive benefit of an organization and its customers. The self-service management interface is still in place while the IT infrastructure resources being collected are internal. In a hybrid cloud environment, external services are leveraged to expand or supplement an internal cloud. Diagrammatically, the three (3) types of cloud computing offerings can be depict by Figure 1

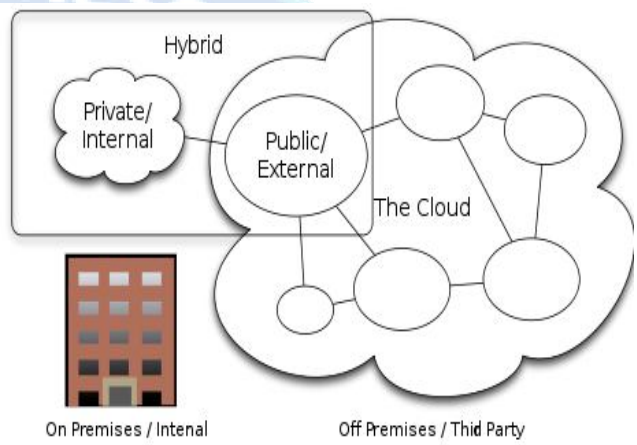

Figure 1: Cloud Computing

3. Literature Review

Salvatore A at. El. The authors discuss the collection and analysis of massive data describing the connections between participants to online social networks. Alternative approaches to social network data collection are defined and evaluated in practice, against the popular Facebook Web site. They propose a privacy-compliant crawlers, two large samples, comprising millions of connections the data is anonymous and organized as an undirected graph. They describe a set of tools that has been developed to analyze specific properties of such social-network graphs, i.e., among others, degree distribution, centrality measures, scaling laws and distribution of friendship. OSNs are among the most intriguing phenomena of the last few years. Data relative to users and relationships among users are not publicly accessible, so they resorted to exploit some techniques derived from Web Data Extraction in order to extract a significant sample of users and relations. They tackled the problem of data extraction by using concepts typical of the graph theory, namely users were represented by nodes of a graph and reflections among users were represented by edges. Two different sampling techniques, namely BFS and Uniform, were adopted in order to explore the graph of OSN, since BFS visiting algorithm is known to introduce a bias in case of an incomplete visit. Even if incomplete for practical limitations, their samples confirm those results related to degree distribution, diameter, clustering coefficient and eigenvalues distribution.

Emilio Ferrara the author introduced Understanding social dynamics that govern human phenomena, such as communications and social relationships is a major problem in current computational social sciences. In particular, given

the unprecedented success of online social networks (OSNs), in his work he is concerned with the analysis of aggregation patterns and social dynamics occurring among users of the largest OSN as the date. In detail, he discuss the mesoscopic features of the community structure of this network, considering the perspective of the communities, which has not yet been studied on such a large scale. To this purpose, he acquired a sample of this network containing millions of users and their social relationships; then, we unveiled the communities representing the aggregation units among which users gather and interact; finally, he analyzed the statistical features of such a network of communities, discovering and characterizing some specific organization patterns followed by individuals interacting in online social networks, that emerge considering different sampling techniques and clustering methodologies. His study provides some clues of the tendency of individuals to establish social interactions in online social networks that eventually contribute to building a well-connected social structure.

Alan Mislove at. el. The authors worked on Online social networking sites like Orkut, YouTube, and Flickr, the most popular sites on the Internet. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. The popularity of these sites provides an opportunity to study the characteristics of online social network graphs at large scale. Understanding these graphs is important, both to improve current systems and to design new applications of online social networks. In their work the authors presents a large-scale measurement study and analysis of the structure of multiple online social networks. They examine data gathered from four popular online social networks: Flickr, YouTube, Live Journal, and Orkut. They crawled the publicly accessible user links on each site, obtaining a large portion of each social network's graph. The data set contains over 11.3 million users and 328 million links. They believed that this is the first study to examine multiple online social networks at scale. Their results confirm the power-law, small-world, and scalefree properties of online social networks. They observe that the indegree of user nodes tends to match the outdegree; that the networks contain a densely connected core of high-degree nodes; and that this core links small groups of strongly clustered, low-degree nodes at the fringes of the network. Finally, we discuss the implications of these structural properties for the design of social network based systems. They presented an analysis of the structural properties of online social networks using data sets collected from four popular sites. Their data shows that social networks are structurally different from previously studied networks, in particular the Web. Social networks have a much higher fraction of symmetric links and also exhibit much higher levels of local clustering. They have outlined how these properties may affect algorithms and applications designed for social networks. They have focused exclusively on the user graph of social networking sites; many of these sites allow users to host content, which in turn can be linked to other users and content. Establishing the structure and dynamics of the content graph is an open problem, the solution to which will enable us to understand how content is introduced in these systems, how data gains popularity, how users interact with popular versus personal data, and so on.

Daqing Zhang and Bin Guo, Zhiwen Yu The authors state that social and community intelligence research aims to reveal individual and group behaviors, social interactions, and community dynamics by mining the digital traces that people leave while interacting with Web applications, static infrastructure, and mobile and wearable devices. The authors specify that SCI aims to identify a set of characteristics or behaviors associated with a social community. Social communities form flexibly from people in the same organization, at the same places, with the same behaviors and interests, and so on, depending on social application requirements. By pooling individual behavior traces and mining the underlying social patterns, an SCI system can extract various social or group behaviors. The extracted social context can be an event such as an open concert, a behavior pattern in daily activity, a relationship within a group, or a significant location. The thrust of SCI pattern mining is to identify user similarity in these social patterns to facilitate offering socially aware services. Unsupervised learning techniques, such as clustering, latent semantic analysis, and matrix factorization, are possible ways to mine social context according to individual behavioral similarities. The process includes the mining and discovery of common social contexts, such as personal characteristics, cuisine preferences, and eagerness to participate socially. It also includes the discovery of undefined social patterns for interest matching and ranking social choices. To enable systems to infer social events on the basis of user context traces, data mining and inference research should aim to bridge the semantic gap between the low level individual activities and high-level social events. The authors believe that SCI represents a new interdisciplinary research and application field and that its scope will continue to expand with innovative applications in the near future. As an emerging research area, SCI still faces challenges, but their resolution will pave the way for new research opportunities. Although existing SCI practices involve only one data source type—Web applications and Internet services, static sensor infrastructure, or mobile and wearable devices—the authors expect to see the rapid growth of research on using the aggregated power of three information sources as well as on enabling innovative SCI-enabled applications.

George Pallis, Demetrios Zeinalipour-Yazti, and Marios D. Dikaiakos The authors state that rapid proliferation of *Online Social Network (OSN)* sites has made a profound impact on the WWW, which tends to reshape its structure, design, and utility. Industry experts believe that OSNs create a potentially transformational change in consumer behavior and will bring a far-reaching impact on traditional industries of content, media, and communications. The work of authors starts by presenting the current status of OSNs through a taxonomy which delineates the spectrum of attributes that relate to these systems. they also presents an overall reference system architecture that aims at capturing the building blocks of prominent OSNs. Additionally, they explores the future trends of OSN systems, presents significant research challenges and discusses their societal and business impact.

OSNs are popular infrastructures for information sharing, communication and interaction on the Internet. With over half a billion users, OSNs are nowadays a mainstream research topic of interest for computer scientists, economists, sociologists etc. the researchers have analyzed and categorized the infrastructural and technical attributes of OSNs. They developed a comprehensive taxonomy for OSNs based on their scope, data model, system model and network model. Authors also present the system architecture for OSNs, where the main components of an OSN platform. However, and despite their impressive success, OSN services face significant challenges that need to be addressed in order to improve the end-user experience and to allow for a healthy market expansion in the future. Consequently, content distribution, scalability and privacy issues are gaining more attention in order to meet up the new technical and infrastructure requirements of the next generation OSNs. This has led to a paradigm shift from a centralized infrastructure to a decentralized one.

Lada A. Adamic, Eytan Adar The researchers found out that anything from the links and text on a user's homepage to the mailing lists the user subscribes to are reflections of social interactions a user has in the real world. In their work they devise techniques and tools to mine this information in order to extract social networks and the exogenous factors underlying the networks' structure. In an analysis of two data sets, from Stanford University and the Massachusetts Institute of Technology (MIT), we show that some factors are better indicators of social connections than others, and that these indicators vary between user populations. Their techniques provide potential applications in automatically inferring real world connections and discovering, labeling, and characterizing communities. They have shown that personal homepages provide a glimpse into the social structure of university communities. Not only do they reveal to us who knows whom, but they give us a context, whether it be a shared dorm, hobby, or research lab. Obtaining data on social networks can be an expensive and time-consuming process of conducting a series of mail, phone or live interviews. Studying social networks online can give us rich insight into how social bonds are created, but requires little more effort than running a crawler on homepages. In their study they have demonstrated a means of leveraging text, mailing list, and in and out-link information to analyze network structure. The researchers have also characterized specific types of items from each of these categories that act as good indicators (individuals associated with an item tend to link to each other) or bad indicators (items which are too general to be indicative of social connections). Furthermore, because indicators vary between communities, we were able to infer characteristics of the communities themselves. Among the numerous applications of these results is the mining of correlations between groups of people, which can be done simply by looking at co-occurrence in homepages of terms associated with each group. Using these techniques in combination with community discovery algorithms yields labeled clusters of users. Thus, not only is it possible to find communities, but they can describe them in a non-obvious way.

Pasquale De Meo at.el The authors discuss the problem of clustering large complex networks plays a key role in several scientific fields ranging from Biology to Sociology and Computer Science. Many approaches to clustering complex networks are based on the idea of maximizing a network modularity function. Some of these approaches can be classified as global because they exploit knowledge about the whole network topology to find clusters. Other approaches, instead, can be interpreted as local because they require only a partial knowledge of the network topology, e.g., the neighbors of a vertex. Global approaches are able to achieve high values of modularity but they do not scale well on large networks and, therefore, they cannot be applied to analyze on-line social networks like Facebook or YouTube. In contrast, local approaches are fast and scale up to large, real-life networks, at the cost of poorer results than those achieved by local methods. In this article we propose a glocal method to maximizing modularity, i.e., our method uses information at the global level, yet its scalability on large networks is comparable to that of local methods. The authors propose a method is called COmplex Network CLUster DEtection (or, shortly, CONCLUDE.) It works in two stages: in the first stage it uses an information-propagation model, based on random and non-backtracking walks of finite length, to compute the importance of each edge in keeping the network connected (called edge centrality.) Then, edge centrality is used to map network vertices onto points of an Euclidean space and to compute distances between all pairs of connected vertices. In the second stage, CONCLUDE uses the distances computed in the first stage to partition the network into clusters. CONCLUDE is computationally efficient since in the average case its cost is roughly linear in the number of edges of the network. Testing on diverse benchmark datasets shows good results, both in times and quality of the clustering; that is the case for synthetic networks with well-defined clusters as well as for real-world network instances.

Their work presents CONCLUDE, an efficient method for detecting clusters in complex networks which is proven to work well in different domains. An early implementation of this algorithm has been already released and its strengths and performance might be assessed independently by other authors. Our ongoing research efforts focus on adopting CONCLUDE in several contexts. They worked on the emergence of a community structure in large online social networks such as Facebook and the assessment of sociological conjectures that involve finding clusters according to importance of edges, for example the strength of the weak ties theory, and i) enhancing the performance of different state-of-the-art clustering algorithms Finally, a long-term research evaluation of our method is planned, in order to cover different domains of application: for example, the application of CONCLUDE could be promising in the context of So far, our algorithm is able to produce a strong partition of the network, but does not allow partitions to overlap each other. This feature will be instrumental in some contexts in which is meaningful that nodes contemporary belong to different clusters

Jure Leskovec Discussed that Online social media represent a fundamental shift of how information is being produced, transferred and consumed. User generated content in the form of blog posts, comments, and tweets establishes a connection between the producers and the consumers of information. Tracking the pulse of the social media outlets, enables companies to gain feedback and insight in how to improve and market products better. For consumers, the abundance of information and opinions from diverse sources helps them tap into the wisdom of crowds, to aid in making more informed decisions. The researcher investigates techniques for social media modeling, analytics and optimization. Firstly they presented methods for collecting large scale social media data and then discuss techniques for coping with and correcting for the effects arising from missing and incomplete data. They proceed by discussing methods for extracting and tracking information as it spreads among the users. Then they examine methods for extracting temporal patterns by which information popularity grows and fades over time. They specify, how to quantify and maximize the influence of media outlets on the popularity and attention given to particular piece of content, and how to build predictive models of information diffusion and adoption. As the information often spreads through implicit social and information networks we present methods for inferring networks of influence and diffusion. Last, we discuss methods for tracking the flow of sentiment through networks and emergence of polarization.

4. CONCLUSION

In this paper, we studied specific formulation of data mining techniques with cloud computing environment. In data mining we want to find useful patterns with different methodology. The main issue with data mining techniques is that the space required for the item set and there operations are very huge. If we combine data mining techniques with cloud computing environment, then we can rent the space from the cloud providers on demand. This solution can solve the problem of huge space and we can apply data mining techniques without taking any consideration of space. In which we can provide data mining concept in cloud computing environment which provide the elasticity to handle a huge amount of data without any hassles about the storage on pay per basis. Then provide cost estimation which shows better cost for the user in the cloud environment in comparison to the non cloud environment.

REFERENCES

[1] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara ,Giacomo Fiumara , Alessandro Provetti_, 'Crawling Facebook for Social Network Analysis Purposes', 2011 ACM.
[2] Emilio Ferrara* 'A large-scale community structure analysis in Facebook' 2012,SpringerOpen Journal
[3] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Drusche, Bobby Bhattacharjee, 'Measurement and Analysis of Online Social Networks' , 2007 ,ACM
[4] Daqing Zhang and Bin Guo, Zhiwen Yu 'The Emergence of Social and Community Intelligence', 2011, IEEE
[5] George Pallis, Demetrios Zeinalipour-Yazti, and Marios D. Dikaiakos, 'Online Social Networks: Status and Trends' 2011, Springer
[6] Lada A. Adamic, Eytan Adar, 'Friends and neighbors on the Web', 2003, Elsevier
[7] Vincent D. Blondel, Anah´i Gajardo, Maureen Heymans, Pierresenellart¶, and Paul Van Doorenk,' A measure of similarity between graph vertices: applications to synonym extraction and web searching', 2004, arXiv:cs
[8] Pasquale De Meo a, Emilio Ferrara b,a, Giacomo Fiumara a, Alessandro Provetti a,c,, 'Mixing local and global information for community detection in large networks', ,2013, arXiv:cs
[9] Jure Leskovec, 'Social Media Analytics: Tracking, Modeling and Predicting the Flow of Information through Networks', 2011, ACM.
[10] S. Sahay et. al., "Comparative Study of Soft Computing Techniques for Ground Water Level Forecasting in a Hard RockArea" International Journal of Research and Development in Applied Science and Engineering, Volume 4, Issue 1, November 2013.