Scope of Machine Learning in Cloud Computing

Gurmeet Singh (07010150), gurmeet@iitg.ernet.in
Under the guidance of **Dr. Diganta Goswami**
8-11-2010

Abstract

Cloud Computing, the on-demand processing and provisioning of services over the Internet, holds a huge scope of using the concepts of Machine Learning for enhanced performance while rendering the cloud more situation-aware and solving some of the challenges in front of the research community. This paper investigates some problems like automation of cloud operations, sharing of data among web services on same cloud and examining behaviour of users in the cloud, which can leverage on suitable learning algorithms.

1 Introduction

Cloud computing is a recent paradigm shift from the client-server architecture which replaced mainframe computers in early 1980s. For sustaining this shift, embedding intelligence in the cloud is a necessary requirement as its availability on the web to a large number of users calls for its ability to scale largely and provide intelligent service to the customers having varied requirements. This intelligence can allow scalability of cloud resources, enhance its performance and give its users- the cloud clients as well as back end operators- the cloud vendors and partners, a better experience using the computing paradigm. Along these lines, the possibility and the need of using machine learning concepts in cloud computing has been acknowledged by the research community [2, 6, 7].

On-demand resource provisioning to the cloud clients is the main focus of cloud computing, which requires estimation of the amount of resources which will be used by the clients' applications prior to resource allocation and client's request execution. This mandates the need of scaling up and scaling down of resources made available to the customers, to meet the service level agreements on one hand and to save cost and power on the other. Scaling of resources by a human operator, although easy, does not seem to be a good option since the size of the clouds is increasing and cloud is also including multiple web services sharing same infrastructure and even the same data [1]. Hence automation of mechanism to allocate resources becomes a necessity. This automation of resource allocator for the web service needs to take into account performance history, performance problems, service level agreements and resource conservation issues while scaling up or down. Such a system can rely on machine learning techniques to efficiently decide the amount of resources necessary for the service.

Further towards automating the cloud is the automation of user troubleshooting, which needs to solve the issues being faced by users considering the performance delivered to the customer in past, past performance problems, solutions adopted in the past and their effect on the cloud's performance before providing solution for the problem reported. This provides scope of utilizing concepts and ideas of machine learning for taking automation to another level. Further, machine learning can be used to monitor usage patterns and draw appropriate conclusions from them to be used in pertinent situations. Examining user behaviour to understand user requirements and expectations is another area which can leverage on learning algorithms for better customer experience.

The paper is structured as follows. The next section discusses some of the work done to utilize machine learning approach along the lines of system performance optimization, system automation and customer satisfaction. Section 3 describes further scope of machine learning techniques in the field of cloud computing. We summarize the observations in the section on Conclusion.

2 Literature Survey

The concept of using machine learning concepts for predicting and increasing system performance is not new. Work has been done in the past along these lines [11, 12]. Recent trends in the

research community have encouraged the use of machine learning concepts in the field of cloud computing [10] to allow the cloud to make conclusions, decisions and taking action in real world setting without any human intervention. The technical report [2], which comprehensively lists the top obstacles and opportunities of cloud computing, foresees the use of technology leveraging on machine learning techniques to automate datacenter operations and aims at developing an Auto-Scalar that automatically manages many aspects of system like dynamic scaling and automatic reaction to performance and correctness problems, as an end to conserve computing resources and maximize profit.

The automation of cloud control and its operation has also been discussed by authors in [7] which uses modeling, control and analysis techniques from statistics and machine learning to capture complex real life workload-performance relationships and adapt to changing conditions which might invalidate simple models. In [6], the proposed architecture: Scalable Consistency Adjustable Data Storage (SCADS) uses machine learning to anticipate performance problems, add and remove resources to meet service level agreements efficiently without downtime, and discusses the possibility of the use of machine learning performance models to provide guidance about possible implementation costs.

Apache Mahout [8] is a scalable machine learning library supporting reasonably large data sets. It is implemented on top of Apache Hadoop which is being used by large customer pool including Facebook, Google and others [9]. Facebook in particular, which has one of the largest known Hadoop storage clusters in the world, uses it as a source of analytics and machine learning. Towards this end, Mahout supports four types of machine learning and data mining algorithms atop Hadoop: Recommendation mining to take users' behaviour and anticipate what users might like, Clustering to take files etc. and group them into sets of related documents, Classification to learn from existing categorized documents about what documents of a particular category look like and assign unlabelled documents to the correct category and Frequent itemset mining or Frequent pattern mining to take a set of item groups and identify which individual items usually appear together.

Although possibility of the use of machine learning techniques for automatic scaling in cloud environment has been mentioned in the literature, other areas where machine learning can contribute remain largely non-investigated. The following section lists some of the opportunities of cloud computing towards this interdisciplinary work of systems and machine learning.

3 Further Scope

This technical report is a result of studying some of the recent developments in the field of Cloud Computing and investigating them for the scope of using machine learning as a tool to further the work been done or as a better alternative to solve the issues in front of the community. Some of the areas which offer opportunity for such interdisciplinary work are related to the issues of hosting web services on the cloud, data sharing in the cloud among various related web services and effective user troubleshooting to aide cloud users in resolving problems faced by them. These issues and their possible solutions using machine learning have been discussed in details below.

3.1 Hosting Web Services

Since the Cloud delivers a hosting environment that does not limit an application to a specific set of resources, Cloud hosting customers do not need to worry about buying new hardware to meet increasing traffic demands or huge traffic spikes. But cloud hosting of the web services which were built independent of cloud environment at the first place poses its own challenges.

CiteSeer^x [13] is a digital library and search engine which disseminates scientific information and literature over the web. Its service oriented interface and loosely coupled architecture offering extensibility and scalability, allow hosting of whole or parts of its framework in a cloud infrastructure [4]. While CiteSeer^x can take advantage of infrastructure on demand, reduced maintenance and elasticity, it needs to face the challenges of the rate of growth of digital information, introduction

of new features and usage. Usage is the most prominent challenge in placing CiteSeer^x into current cloud infrastructure offerings [5].

Monitoring usage patterns and taking required actions to meet SLAs is a crucial requirement of cloud offerings. This can be met using the proposed learning aspect of cloud. Monitoring done over the past usage scenario help predict future usage patters as well as anticipate possible resource needs of the service. Following the cloud automation and statistical machine learning techniques adopted by authors in [2, 6, 7] it is possible to provide usage monitoring and handling peek usage scenario efficiently in cloud hosted web services.

Scope of machine learning in cloud hosted services is not just limited to usage monitoring, but also extends to other areas like analysing user behaviour. One of the goals of CiteSeer^x is to examine the behaviour of user in the cloud [4] and try to adapt the system according to changing behavioural patterns. As mentioned earlier, work on examining behaviour of users has been done in cloud environment by Apache in building a machine learning library Mahout which not only examines the user behaviour but also anticipates what they might like using Recommendation mining. Such, or similar work can possibly be done in the field of hosting web services on cloud which need analogous user behaviour monitoring techniques.

3.2 Data Sharing Among Web Services

With the advent of its growing size, the cloud is increasingly becoming a host for multiple web services, rather than each web service opting for a traditional in-house datacenter. This holds opportunity of convenient, efficient and large scale data sharing among web services located in the same cloud [1]. A major challenge in this context is performance isolation of more than one web services which might get requests for the shared data objects leading to one service blocking the requests of the other, although temporarily.

The paper on CloudViews [1] uses views of databases to allow services to create and share views of their data over the common storage infrastructure. A web service can hence access the data of another web service which has shared a view of its data with the former. This database style view abstraction supports flexible, secure, efficient and performance isolated sharing of data. The isolation stems from the fact that in CloudViews scheduling is done fairly for the views, each one of which has its own queue. The fair scheduling algorithm for the queues is suggested by the authors and is presently under investigation.

While the queues for each view will have fair scheduling, preventing the performance problems of one web service blocking the data for another, the system still relies on the same data objects which are sequentially or alternately, although fairly, accessed by multiple web services. Alternatively, a better option can be to create a copy of the data being accessed by more than one web services, to facilitate parallel access to the services rather than alternate one. When a Secondary web service having permission to view data owned by a Primary web service, requests the data objects which are also needed by the Primary, then the cloud can copy a part of Primary's data for the Secondary. While Secondary accesses these data objects, cloud can create copy of additional data which is likely to be accessed in the near future, depending on the past data-access-patterns of Primary's data objects by the Secondary, learnt using an efficient machine learning algorithm.

The access patterns can be broadly of two kinds, first is linear access in which case more and more data must be copied, and second is repeated or circular access where same copy of the data will be accessed and hence needs to be updated at the original location if changed to achieve consistency. For the prior case, cloud needs to learn the amount of time for which the copy of a particular data object needs to be kept, and for the latter, cloud should timely update of the original copy of the data if it has been modified by the Secondary and vice versa. Machine learning algorithms, in conjunction with replica management techniques can be used to isolate the performance of web services sharing the same data independently while depending on the same cloud infrastructure for fast and cheap sharing.

3.3 User Troubleshooting

The idea of service level agreements and user satisfaction is central to the domain of cloud computing. Helping the users to resolve the problems faced by them is mandatory to please the customers as well as provide a good service. Towards this end, [3] investigates the problems reported by users and the methods deployed by the vendors to solve them. It also proposes a set of three principles for designing Infrastructure-as-a-Service support models for better user troubleshooting: developing tools targeted at debugging new features, developing techniques to automate operator task, and providing a vehicle to gather and transfer information between operator and user. While the first principle can be achieved by suitable debugging software, the latter two principles can efficiently rely on algorithms and techniques of machine learning and artificial intelligence.

Automating the tasks of human operator for user troubleshooting involves understanding the problem being reported and solving the problem considering the past performance of cloud as seen by the customer, problems encountered in the past, solutions applied in past for similar problems and their effect on cloud's performance. It also includes taking appropriate steps to resolve the issue meeting the desired SLAs and observing if the problem still persists. This can be done by modeling the human operator in the limited setting of cloud environment and decision making regarding cloud problems. Several attempts have been made in the past to model a human operator in restricted environment for efficient automation of human decision making. For instance, modeling the decision making of a human operator in automated manufacturing systems is discussed in [14], and several other related works in different settings are also available.

4 Conclusion

Cloud computing has immense scope of using techniques of machine learning to enhance performance and increase customer satisfaction and operator experience. Learning algorithms can be utilized to automate the cloud operations like scaling the resources and user troubleshooting, to monitor usage patterns and taking necessary actions to meet SLAs, to provide efficient mechanism of data sharing among more than one web services hosted on a public cloud and to examine behaviour of users for understanding their requirements and expectations from cloud.

Machine learning is a vast field and has diverse applications in systems. Further investigation into cloud computing literature can reveal more such areas which can effectively leverage on machine learning and related topics.

5 Acknowledgements

I would like to thank my guide Dr. Diganta Goswami for encouraging work on the topic of my interest, sharing his knowledge with me and supporting me throughout the term paper. I would also like to extend thanks to Dr. Vijaya Saradi for taking out time to discuss the scope of machine learning held by cloud computing and making me acquainted with the basics of learning algorithms used in systems.

References

[1] R. Geambasu, S. D. Gribble, and H. M. Levy. "CloudViews: communal data sharing in public clouds". In *HotCloud 09 Workshop on Hot Topics in Cloud Computing*, 2009.

[2] Armbrust M., Fox A., Griffith R., Joseph A.D., Katz R., Konwinski A., Lee H., Patterson D., Rabkin A., Stoica I. Zaharia M. "Above the Clouds: A Berkeley View of Cloud Computing". In *Technical Report No. UCB/EECS-2009-28, UC Berkeley Reliable Adaptive Distributed Systems Laboratory*, 2009.

[3] Theophilus Benson, Sambit Sahu, Aditya Akella, and Anees Shaikh. "A First Look at Problems in the Cloud". In *HotCloud'10*, Boston, MA, 2010.

[4] Pradeep Teregowda, Bhuvan Urgaonkar, and C. Lee Giles. "CiteSeerX: A Cloud Perspective". In *Proceedings of the Second USENIX Workshop on Hot Topics in Cloud Computing*, 2010.

[5] Pradeep Teregowda, Bhuvan Urgaonkar, and C. Lee Giles. "Cloud Computing: A Digital Libraries Perspective". In *P IEEE Third International Conference on Cloud Computing (CLOUD 2010)*, Miami, FL, July 2010.

[6] Michael Armbrust, Armando Fox, David A. Patterson, Nick Lanham, Beth Trushkowsky, Jesse Trutna, and Haruki Oh. "SCADS: Scale-Independent Storage for Social Computing Applications". In *4th Biennial Conference on Innovative Data Systems Research (CIDR)*, January 4-7, 2009, Asilomar, California, USA.

[7] Peter Bod, Rean Griffith, Charles Sutton, Armando Fox, Michael Jordan, David Patterson. "Statistical Machine Learning Makes Automatic Control Practical for Internet Datacenters". In *HotCloud 2009: Proceedings of the Workshop on Hot Topics in Cloud Computing*, San Diego, CA, 2009.

[8] "Mahout". *http://mahout.apache.org* .

[9] "Hadoop Wiki". *http://wiki.apache.org/hadoop/PoweredBy* .

[10] "Steve Ballmer: Cloud Computing". *http://www.microsoft.com/presspass/exec/steve/2010/03-04Cloud.mspx* .

[11] Archana Ganapathi, Harumi Kuno, Umeshwar Dayal, Janet L. Wiener, Armando Fox, Michael Jordan, David Patterson. "Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning". In *Proc International Conference on Data Engineering*, 2009.

[12] Archana Sulochana Ganapathi. "Predicting and Optimizing System Utilization and Performance via Statistical Machine Learning". In *Technical Report No. UCB/EECS-2009-181*, 2009.

[13] "CiteSeer^x". http://citeseerx.ist.psu.edu/ .

[14] Xiaobing Zhao, Young-Jun Son. "BDI-based Human Decision-Making Model in Automated Manufacturing Systems". In *International Journal of Modeling and Simulation*, 28(3), 2008, 347-356..