EMERGING PARADIGMS FOR ANALYZING, PROCESSING AND MAKING SENSE OF LARGE HETEROGENEOUS DATASETS.

L K Sravanthi Potti[1], Dr. Satheesh Kumar Nagineni[2]
Department of Computer Science and Engineering
[1,2]OPJS University, Churu, Rajasthan

Abstract

This thesis explores the problem of large scale Web mining by using Data Intensive Scalable Computing (DISC) systems. Web mining aims to extract useful information and models from data on the Web, the largest repository ever created. DISC systems are an emerging technology for processing huge datasets in parallel on large computer clusters.

Challenges arise from both themes of research. The Web is heterogeneous: data lives in various formats that are best modeled in different ways. Effectively extracting information requires careful design of algorithms for specific categories of data. The Web is huge, but DISC systems offer a platform for building scalable solutions. However, they provide restricted computing primitives for the sake of performance. Efficiently harnessing the power of parallelism offered by DISC systems involves rethinking traditional algorithms.

1. Introduction

An incredible “data deluge” is currently drowning the world. Data sources are everywhere, from Web 2.0 and user-generated content to large scientific experiments, from social networks to wireless sensor networks. This massive amount of data is a valuable asset in our information society.Data analysis is the process of inspecting data in order to extract useful information. Decision makers commonly use this information to drive their choices. The quality of the information extracted by this process greatly benefits from the availability of extensive datasets. The Web is the biggest and fastest growing data repository in the world. Its size and diversity make it the ideal resource to mine for useful information. Data on the Web is very diverse in both content and format. Consequently, algorithms for Web mining need to take into account the specific characteristics of the data to be efficient.

As we enter the “petabyte age”, traditional approaches for data analy-sis begin to show their limits. Commonly available data analysis tools are unable to keep up with the increase in size, diversity and rate of change of the Web. Data Intensive Scalable Computing is an emerging alternative technology for large scale data analysis. DISC systems combine both storage and computing in a distributed and virtualized manner. These systems are built to scale to thousands of computers, and focus on fault tolerance, cost effectiveness and ease of use.

1.1 The Data Deluge

How would you sort 1GB of data? Today's computers have enough memory to keep this quantity of data, so any optimal in-memory al-gorithm will suffice. What if you had to sort 100 GB of data? Even if systems with more than 100 GB of memory exist, they are by no means common or cheap. So the best solution is to use a disk based sorting algorithm. However, what if you had 10 TB of data to sort? At a transfer rate of about 100 MB/s for a normal disk it would take more than one day to make a single pass over the dataset. In this case the bandwidth between memory and disk is the bottleneck. In any case, today's disks are usually 1 to 2 TB in size, which means that just to hold the data we need multiple disks. In order to obtain acceptable completion times, we also need to use multiple computers and a parallel algorithm.

We can think that the intrinsic characteristics of the object to be an-alyzed demand modifications to traditional data managing procedures.

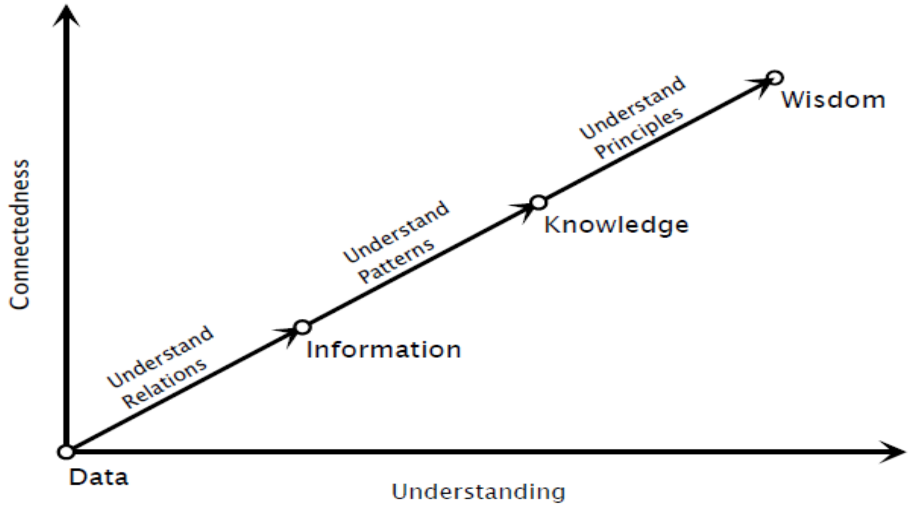

Figure 1: Data Information Knowledge Wisdom hierarchy.

Alternatively, we can take the point of view of the subject who needs to manage the data. The emphasis is thus on user requirements such as throughput and latency. In either case, all the previous definitions hint to the fact that big data is a driver for research.

1.2 Mining the Web

The Web is easily the single largest publicly accessible data source in the world (Liu, 2007). The continuous usage of the Web has accelerated its growth. People and companies keep adding to the already enormous mass of pages already present.

In the last decade the Web has increased its importance to the point of becoming the center of our digital lives (Hammersley, 2011). People shop and read news on the Web, governments offer public services through it and enterprises develop Web marketing strategies. Investments in Web advertising have surpassed the ones in television and newspaper in most countries. This is a clear testament to the importance of the Web.

The estimated size of the indexable Web was at least 11.5 billion pages as of January 2005 (Gulli and Signorini, 2005). Today, the Web size is estimated between 50 and 100 billion pages and roughly doubling every eight months (Baeza-Yates and Ribeiro-Neto, 2011), faster than Moore's law. Furthermore, the Web has become infinite for practical purpose, as it is possible to generate an infinite number of dynamic pages. As a result, there is on the Web an abundance of data with growing value.

Web structure mining mines the hyperlink structure of the Web usinggraph theory. For example, links are used by search engines to find important Web pages, or in social networks to discover communi-ties of users who share common interests.

Web content mining analyzes Web page contents. Web content miningdiffers from traditional data and text mining mainly because of the semi-structured and multimedial nature of Web pages. For example, it is possible to automatically classify and cluster Web pages according to their topics but it is also possible to mine customer product reviews to discover consumer sentiments.

Web usage mining extracts information from user access patterns foundin Web server logs, which record the pages visited by each user, and from search patterns found in query logs, which record the terms searched by each user. Web usage mining investigates what users are interested in on the Web.

Mining the Web is typically deemed highly promising and rewarding. However, it is by no means an easy task and there is a flip side of the coin: data found on the Web is extremely noisy.

2. Taxonomy of Web data

The Web is a very diverse place. It is an open platform where anybody can add his own contribution. Resultingly, information on the Web is heterogeneous. Almost any kind of information can be found on it, usually reproduced in a proliferation of different formats. As a consequence, the categories of data available on the Web are quite varied.

Data of all kinds exist on the Web: semi-structured Web pages, structured tables, unstructured texts, explicit and implicit links, and multimedia files (images, audios, and videos) just to name a few. A complete classification of the categories of data on the Web is out of the scope of this thesis. However, we present next what we consider to be the most common and representative categories, the ones on which we focus our attention. Most of the Web fits one of these three categories:

Bags are unordered collections of items. The Web can be seen as a col-lections of documents when ignoring hyperlinks. Web sites that collect one specific kind of items (e.g. flickr or YouTube) can also be modeled as bags. The items in the bag are typically represented as sets, multisets or vectors. Most classical problems like similarity, clustering and frequent itemset mining are defined over bags.

Graphs are defined by a set of vertexes connected by a set of edges. TheWeb link structure and social networks fit in this category. Graph are an extremely flexible data model as almost anything can be seen as a graph. They can also be generated from predicates on a set of items (e.g. similarity graph, query flow graph). Graph algorithms like PageRank, community detection and matching are commonly employed to solve problems in Web and social network mining.

Streams are unbounded sequences of items ordered by time. Search queries and click streams are traditional examples, but streams are generated as well by news portals, micro-blogging services and real-time Web sites like twitter and “status updates” on social net-works like Facebook, Google+ and LinkedIn. Differently from time series, Web streams are textual, multimedial or have rich metadata. Traditional stream mining problems are clustering, classification and estimation of frequency moments.

3. Data Intensive Scalable Computing

Let us highlight some of the requirements for a system used to perform data intensive computing on large datasets. Given the effort to find a novel solution and the fact that data sizes are ever growing, this solution should be applicable for a long period of time. Thus the most important requirement a solution has to satisfy is scalability.

Scalability is defined as “the ability of a system to accept increased input volume without impacting the profits”. This means that the gains from the input increment should be proportional to the increment itself. This is a broad definition used also in other fields like economy. For a system to be fully scalable, the size of its input should not be a design parameter. Forcing the system designer to take into account all possible deployment sizes in order to cope with different input sizes leads to a scalable architecture without fundamental bottlenecks.

However, apart from scalability, there are other requirements for a large scale data intensive computing system. Real world systems cost money to build and operate. Companies attempt to find the most cost effective way of building a large system because it usually requires a sig-nificant money investment. Partial upgradability is an important money saving feature, and is more easily attained with a loosely coupled sys-tem. Operational costs like system administrators’ salaries account for a large share of the budget of IT departments. To be profitable, large scale systems must require as little human intervention as possible. Therefore autonomic systems are preferable, systems that are self-configuring, self-tuning and self-healing. In this respect fault tolerance is a key property.

Fault tolerance is “the property of a system to operate properly in spite of the failure of some of its components”. When dealing with a large number of systems, the probability that a disk breaks or a server crashes raises dramatically: it is the norm rather than the exception. A performance degradation is acceptable as long as the systems does not halt completely. A denial of service of a system has a negative economic impact, especially for Web-based companies. The goal of fault tolerance techniques is to create a highly available system.

4. Contributions

DISC systems are an emerging technology in the data analysis field that can be used to capitalize on massive datasets coming from the Web. “There is no data like more data” is a famous motto that epitomizes the opportunity to extract significant information by exploiting very large volumes of data. Information represents a competitive advantage for actors operating in the information society, an advantage that is all the greater the sooner it is achieved. Therefore, in the limit online analytics will become an invaluable support for decision making.

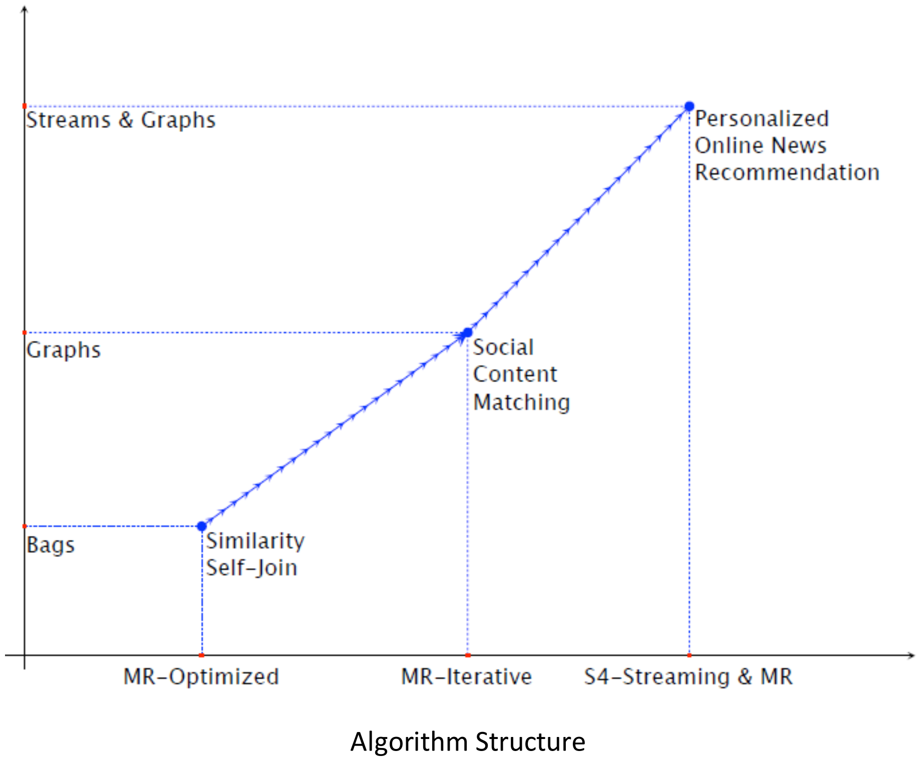

Figure 2: Complexity of contributed algorithms.

5. Conclusion

Data Intensive Scalable Computing is an emerging technology in the data analysis field that can be used to capitalize on massive datasets coming from the Web. In particular MapReduce and streaming are the two most promising emerging paradigms for analyzing, processing and making sense of large heterogeneous datasets. While MapReduce offers the capability to analyze large batches of stored data, streaming solutions offer the ability to continuously process unbounded streams of data online.

“It is not who has the best algorithms that wins, it is who has more data” is a well-known aphorism. The importance of big data is widely recognized both in academia and in industry, and the Web is the largest public data repository available. Information extracted from the Web has the potential to have a big impact in our society.

6. References

1. Fabian Abel, Qi Gao, G.J. Houben, and Ke Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In UMAP ’11: 19th Inter-national Conference on User Modeling, Adaption and Personalization, pages 1–12. Springer, 2011. 106

2. G Adomavicius and ATuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans-actions on Knowledge and Data Engineering, pages 734–749, 2005. 99
3. Foto N Afrati and Jeffrey D. Ullman. A New Computation Model for Cluster Computing. 2009. URL http://ilpubs.stanford.edu:8090/953/. 27,53
4. Foto N. Afrati and Jeffrey D. Ullman. Optimizing Joins in a Map-Reduce En-vironment. In EDBT ’10: 13th International Conference on Extending Database Technology, pages 99—-110, 2010. 34
5. Foto N. Afrati and Jeffrey D. Ullman.Optimizing Multiway Joins in a Map-Reduce Environment. IEEE Transactions on Knowledge and Data Engineering, 23(9):1282 – 1298, 2011. 34
6. Foto N. Afrati, Anish Das Sarma, David Menestrina, AdityaParameswaran, and Jeffrey D. Ullman. Fuzzy Joins Using MapReduce. Technical report, Stanford InfoLab, July 2011a. URL http://ilpubs.stanford.edu:8090/1006. 34
7. Foto N. Afrati, DimitrisFotakis, and Jeffrey D. Ullman. Enumerating Subgraph Instances Using Map-Reduce. Technical report, Stanford University, 2011b. URL http://ilpubs.stanford.edu:8090/1020/. 35
8. Charu C. Aggarwal. Data Streams: Models and Algorithms. Springer, 2007. ISBN 0387287590. 35.