Mining Big Data in Real Time

Albert Bifet
Yahoo! Research Barcelona
Avinguda Diagonal 177, 8th floor
Barcelona, 08018, Catalonia, Spain
E-mail: abifet@yahoo-inc.com

Keywords: big data, data streams, data mining

Received: December 15, 2012

Streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge from what is happening now, allowing organizations to react quickly when problems appear or to detect new trends helping to improve their performance. Evolving data streams are contributing to the growth of data created over the last few years. We are creating the same quantity of data every two days, as we created from the dawn of time up until 2003. Evolving data streams methods are becoming a low-cost, green methodology for real time online prediction and analysis. We discuss the current and future trends of mining evolving data streams, and the challenges that the field will have to overcome during the next years.

Povzetek: Prispevek opisuje rudarjenje velikih količin podatkov na osnovi pretakanja podatkov v podjetjih.

1 Introduction

Nowadays, the quantity of data that is created every two days is estimated to be 5 exabytes. This amount of data is similar to the amount of data created from the dawn of time up until 2003. Moreover, it was estimated that 2007 was the first year in which it was not possible to store all the data that we are producing. This massive amount of data opens new challenging discovery tasks.

Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring and traffic management, log records or click-streams in web exploring, manufacturing processes, call detail records, email, blogging, twitter posts and others [17]. In fact, all data generated can be considered as streaming data or as a snapshot of streaming data, since it is obtained from an interval of time.

In the data stream model, data arrive at high speed, and algorithms that process them must do so under very strict constraints of space and time. Consequently, data streams pose several challenges for data mining algorithm design. First, algorithms must make use of limited resources (time and memory). Second, they must deal with data whose nature or distribution changes over time.

We need to deal with resources in an efficient and low-cost way. *Green computing* is the study and practice of using computing resources efficiently. A main approach to green computing is based on algorithmic efficiency. In data stream mining, we are interested in three main dimensions:

- accuracy
- amount of space (computer memory) necessary
- the time required to learn from training examples and to predict

These dimensions are typically interdependent: adjusting the time and space used by an algorithm can influence accuracy. By storing more pre-computed information, such as look up tables, an algorithm can run faster at the expense of space. An algorithm can also run faster by processing less information, either by stopping early or storing less, thus having less data to process. The more time an algorithm has, the more likely it is that accuracy can be increased.

The issue of the measurement of three evaluation dimensions simultaneously has led to another important issue in data stream mining, namely estimating the combined cost of performing the learning and prediction processes in terms of time and memory. As an example, several rental cost options exist:

- Cost per hour of usage: Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. Cost depends on the time and on the machine rented (small instance with 1.7 GB, large with 7.5 GB or extra large with 15 GB).
- Cost per hour and memory used: GoGrid is a web service similar to Amazon EC2, but it charges by RAM-Hours. Every GB of RAM deployed for 1 hour equals one RAM-Hour.

In [11, 9] the use of RAM-Hours was introduced as an evaluation measure of the resources used by streaming algorithms. Every GB of RAM deployed for 1 hour equals one RAM-Hour.

The structure of this paper is as follows: Section 2 introduces Big Data, Section 3 shows some open source tools, Sections 5, 6, and 7 discuss future trends and introduce a new problem, application and technique for real time Big Data mining, and finally Section 8 concludes the paper.

2 Big data

Big Data is a new term used to identify the datasets that due to their large size, we can not manage them with the typical data mining software tools. Instead of defining "Big Data" as datasets of a concrete large size, for example in the order of magnitude of petabytes, the definition is related to the fact that the dataset is too big to be managed without using new algorithms or technologies.

The McKinsey Global Institute (MGI) published a report on Big Data [26] that describes the business oportunities that big data opens: a potential value of $300 billion in the US health care, $149 billion in European governement administration or improving the operating margin of retailer companies by 60 percent.

Big Data analytics is becoming an important tool to improve efficiency and quality in organizations, and its importance is going to increase in the next years.

There are two main strategies for dealing with big data: sampling and using distributed systems. Sampling is based in the fact that if the dataset is too large and we can not use all the examples, we can obtain an approximate solution using a subset of the examples. A good sampling method will try to select the best instances, to have a good performance using a small quantity of memory and time.

An alternative to sampling is the use of probabilistic techniques. Backstrom, Boldi, Rosa, Ugander and Vigna [7] computed the average distance of friendship links between users in Facebook. They repeated the experiment that Stanley Milgram did in 1967 [28], where he challenged people to send postcards to specific people around US using only direct acquaintances. Milgram obtained a number between 4.4 and 5.7, so the notion of six degrees of separation was confirmed. The experiments using Facebook showed a four degrees of separation pattern. To run these experiments, these researchers used HyperANF, a software tool by Boldi, Rosa and Vigna [13] that improved ANF. ANF is a fast and scalable tool for data mining in massive graphs [30] that computes approximations to the neighbourhood function of nodes in massive graphs. The *neighbourhood function* of a node n and distance h is defined as the number of nodes at a certain distance h reachable from node n. This function is computed using the set of nodes that can be reachable from a distance $h-1$ using probabilistic counter data structures.

The number of users of Facebook is more than 800 million users, but they managed to compute the average distance between two users on Facebook only using one machine. As one of the authors of this paper, Paolo Boldi, once said "Big data does not need big machines, it needs big intelligence".

The most popular distributed systems used nowadays are based in the *map-reduce* framework. The map-reduce methodology started in Google, as a way to perform crawling of the web in a faster way. Hadoop is a open-source implementation of map-reduce started in Yahoo! and is being used in many non-streaming big data analysis.

The map-reduce model divides algorithms in two main steps: map and reduce, inspired in ideas in functional programming. The input data is split into several datasets and each split is send to a mapper, that will transform the data. The output of the mappers will be combined in reducers, that will produce the final output of the algorithm.

Nowadays, in business, more than size and scale, what it is important is speed and agility. As David Meerman Scott explains in his book "Real-Time Marketing & PR" [33], it is important for companies to know what is happening right now, in real time, to be able to react, and more important, to anticipate and detect new business opportunities.

Finally, we would like to mention the work that Global Pulse is doing [37] using Big Data to improve life in developing countries. Global Pulse is a United Nations initiative, launched in 2009, that functions as an innovative lab, and that is based in mining Big Data for developing countries. They pursue a strategy that consists of 1) researching innovative methods and techniques for analyzing real-time digital data to detect early emerging vulnerabilities; 2) assembling free and open source technology toolkit for analyzing real-time data and sharing hypotheses; and 3) establishing an integrated, global network of Pulse Labs, to pilot the approach at country level.

Global Pulse describe the main opportunities Big Data offers to developing countries in their White paper "Big Data for Development: Challenges & Opportunities"[22]:

- Early warning: develop fast response in time of crisis, detecting anomalies in the usage of digital media
- Real-time awareness: design programs and policies with a more fine-grained representation of reality
- Real-time feedback: check what policies and programs fails, monitoring it in real time, and using this feedback make the needed changes

The Big Data mining revolution is not restricted to the industrialized world, as mobiles are spreading in developing countries as well. It is estimated than there are over five billion mobile phones, and that 80% are located in developing countries.

3 Tools: open source revolution

The Big Data phenomenon is intrinsically related to the open source software. Large companies as Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as:

- Apache Hadoop [3]: software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.

- Apache Pig [6]: software for analyzing large data sets that consists of a high-level language similar to SQL for expressing data analysis programs, coupled with infrastructure for evaluating these programs. It contains a compiler that produces sequences of MapReduce programs.

- Cascading [15]: software abstraction layer for Hadoop, intended to hide the underlying complexity of MapReduce jobs. Cascading allows users to create and execute data processing workflows on Hadoop clusters using any JVM-based language.

- Scribe [16]: server software developed by Facebook and released in 2008. It is intended for aggregating log data streamed in real time from a large number of servers.

- Apache HBase [4]: non-relational columnar distributed database designed to run on top of Hadoop Distributed Filesystem (HDFS). It is written in Java and modeled after Google's BigTable. HBase is an example if a NoSQL data store.

- Apache Cassandra [2]: another open source distributed database management system developed by Facebook. Cassandra is used by Netflix, which uses Cassandra as the back-end database for its streaming services.

- Apache S4 [29]: platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.

- Storm [34]: software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter.

In Big Data Mining, there are many open source initiatives. The most popular are the following:

- Apache Mahout [5]: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.

- MOA [9]: Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The `streams` framework [12] provides an environment for defining and running stream processes using simple XML based definitions and is able to use MOA.

- R [32]: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.

- Vowpal Wabbit [21]: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning.

- PEGASUS [19]: big graph mining system built on top of MAPREDUCE. It allows to find patterns and anomalies in massive real-world graphs.

- GraphLab [25]: high-level graph-parallel system built without using MAPREDUCE. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices.

4 Big data mining academic and industrial research

There are many interesting ongoing academic and industrial research in the area, published in main conferences such as KDD, ICDM, ECML-PKDD, or journals such as "Data Mining and Knowledge Discovery", "Machine Learning" or "Journal of Machine Learning Research". Some examples are the following, published in the SIGKDD Explorations Issue of December 2012:

- Mining Heterogeneous Information Networks: A Structural Analysis Approach by *Yizhou Sun and Jiawei Han (University of Illinois at Urbana-Champaign)* [35].

- Big Graph Mining: Algorithms and discoveries by *U Kang and Christos Faloutsos (Carnegie Mellon University)* [20].

- Scaling Big Data Mining Infrastructure: The Twitter Experience by *Jimmy Lin and Dmitriy Ryaboy (Twitter,Inc.)* [23].

- Mining Large Streams of User Data for Personalized Recommendations by *Xavier Amatriain (Netflix)* [1].

5 New problems: structured classification

A new important and challenging task may be the structured pattern classification problem. *Patterns* are elements of (possibly infinite) sets endowed with a partial order relation $\preceq$. Examples of patterns are itemsets, sequences, trees and graphs.

The structured pattern classification problem is defined as follows. A set of examples of the form (t, y) is given, where y is a discrete class label and t is a pattern. The goal is to produce from these examples a model $\hat{y} = f(t)$ that will predict the classes y of future pattern examples

Most standard classification methods can only deal with vector data, which is but one of many possible pattern structures. To apply them to other types of patterns, such as graphs, we can use the following approach: we convert the pattern classification problem into a vector classification learning task, transforming patterns into vectors of attributes. Each attribute denotes the presence or absence of particular subpatterns, and we create attributes for all frequent subpatterns, or for a subset of these.

As the number of frequent subpatterns may be very large, we may perform a feature selection process, selecting a subset of these frequent subpatterns, maintaining exactly or approximately the same information.

The structured output classification problem is even more challenging and is defined as follows. A set of examples of the form (t, y) is given, where t and y are patterns. The goal is to produce from these examples a pattern model $\hat{y} = f(t)$ that will predict the patterns y of future pattern examples. A way to deal with a structured output classification problem is to convert it to a multi-label classification problem, where the output pattern y is converted into a set of labels representing a subset of its frequents subpatterns.

Therefore, data stream multi-label classification methods may offer a solution to the structured output classification problem.

This problem has been studied in the non-streaming setting using relational learning techniques, and has been well developed within inductive logic programming and statistical relational learning [18].

6 New applications: social networks

A future trend in mining evolving data streams will be how to analyze data from social networks and micro-blogging applications such as Twitter. Micro-blogs and Twitter data follow the data stream model. Twitter data arrive at high speed, and algorithms that process them must do so under very strict constraints of space and time.

The main Twitter data stream that provides all messages from every user in real-time is called Firehose and was made available to developers in 2010. This streaming data opens new challenging knowledge discovery issues. In April 2010, Twitter had 106 million registered users, and 180 million unique visitors every month. New users were signing up at a rate of 300,000 per day. Twitter's search engine received around 600 million search queries per day, and Twitter received a total of 3 billion requests a day via its API. It could not be clearer in this application domain that to deal with this amount and rate of data, streaming techniques are needed.

Sentiment analysis can be cast as a classification problem where the task is to classify messages into two categories depending on whether they convey positive or negative feelings. See [31] for a survey of sentiment analysis, and [24] for opinion mining techniques.

To build classifiers for sentiment analysis, we need to collect training data so that we can apply appropriate learning algorithms. Labeling tweets manually as positive or negative is a laborious and expensive, if not impossible, task. However, a significant advantage of Twitter data is that many tweets have author-provided sentiment indicators: changing sentiment is implicit in the use of various types of emoticons. *Smileys* or *emoticons* are visual cues that are associated with emotional states. They are constructed using the characters available on a standard keyboard, representing a facial expression of emotion. Hence we may use these to label our training data.

When the author of a tweet uses an emoticon, they are annotating their own text with an emotional state. Such annotated tweets can be used to train a sentiment classifier [8, 10].

Another interesting application is the NELL (Never-Ending Language Learner) system developed by the group of Tom Mitchell [14, 36] at Carnegie Mellon University. The goal of this system is to build a never-ending machine learning system that acquires the ability to extract structured information from unstructured web pages. It involves text analysis of 500 million web pages and access to the remainder of the web through search engine APIs. NELL runs 24 hours per day, continuously, to perform two ongoing tasks: extract new instances of categories and relations, and learn to read better than yesterday.

7 New techniques: Hadoop, S4 or Storm

A way to speed up the mining of streaming learners is to distribute the training process onto several machines. Hadoop MapReduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes.

A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.

Apache S4 [29] is a platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time. Storm [34] from Twitter uses a similar approach. Ensemble learning classifiers are easier to scale and parallelize than single classifier methods. They are the first, most obvious, candidate methods to implement using parallel techniques.

It is not clear yet how an optimal architecture for analytics systems may be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [27]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in realtime by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, extensible, allows ad hoc queries, minimal maintenance, and debuggable.

8 Conclusions

We have discussed the challenges that in our opinion, mining evolving data streams will have to deal during the next years. We have outlined new areas for research. These include structured classification and associated application areas as social networks.

Our ability to handle many exabytes of data across many application areas in the future will be crucially dependent on the existence of a rich variety of datasets, techniques and software frameworks. There is no doubt that data stream mining offers many challenges and equally many opportunites as the quantity of data generated in real time is going to continue growing.

References

[1] X. Amatriain. Mining large streams of user data for personalized recommendations. *SIGKDD Explorations*, 14(2), 2012.

[2] Apache Cassandra, http://cassandra. apache.org.

[3] Apache Hadoop, http://hadoop.apache.org.

[4] Apache HBase, http://hbase.apache.org.

[5] Apache Mahout, http://mahout.apache.org.

[6] Apache Pig, http://www.pig.apache.org/.

[7] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four Degrees of Separation. *CoRR*, abs/1111.4570, 2011.

[8] A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter streaming data. In *Proc 13th International Conference on Discovery Science*, Canberra, Australia, pages 1–15. Springer, 2010.

[9] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. *Journal of Machine Learning Research (JMLR)*, 2010.

[10] A. Bifet, G. Holmes, and B. Pfahringer. Moa-tweetreader: Real-time analysis in twitter streaming data. In *Discovery Science*, pages 46–60, 2011.

[11] A. Bifet, G. Holmes, B. Pfahringer, and E. Frank. Fast perceptron decision tree learning from evolving data streams. In *PAKDD*, 2010.

[12] C. Bockermann and H. Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012.

[13] P. Boldi, M. Rosa, and S. Vigna. HyperANF: approximating the neighbourhood function of very large graphs on a budget. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 625–634, 2011.

[14] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[15] Cascading, http://www.cascading.org/.

[16] Facebook Scribe, https://github.com/ facebook/scribe.

[17] J. Gama. *Knowledge discovery from data streams*. Chapman & Hall/CRC, 2010.

[18] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[19] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.

[20] U. Kang and C. Faloutsos. Big graph mining: Algorithms and discoveries. *SIGKDD Explorations*, 14(2), 2012.

[21] J. Langford. Vowpal Wabbit, http://hunch.net/˜vw/, 2011.

[22] E. Letouzé. Big Data for Development: Opportunities & Challenges. May 2011.

[23] J. Lin and D. Ryaboy. Scaling big data mining infrastructure: The twitter experience. *SIGKDD Explorations*, 14(2), 2012.

[24] B. Liu. *Web data mining; Exploring hyperlinks, contents, and usage data*. Springer, 2006.

[25] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, July 2010.

[26] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. May 2011.

[27] N. Marz and J. Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2013.

[28] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[29] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. In *ICDM Workshops*, pages 170–177, 2010.

[30] C. R. Palmer, P. B. Gibbons, and C. Faloutsos. ANF: a fast and scalable tool for data mining in massive graphs. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 81–90, 2002.

[31] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[33] D. M. Scott. *Real-Time Marketing and PR, Revised: How to Instantly Engage Your Market, Connect with Customers, and Create Products that Grow Your Business Now*. Wiley Desktop Editions Series. John Wiley & Sons, 2011.

[34] Storm, http://storm-project.net.

[35] Y. Sun and J. Han. Mining heterogeneous information networks: A structural analysis approach. *SIGKDD Explorations*, 14(2), 2012.

[36] P. P. Talukdar, D. Wijaya, and T. Mitchell. Coupled temporal scoping of relational facts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM)*, Seattle, Washington, USA, February 2012. Association for Computing Machinery.

[37] United Nations Global Pulse, http://www.unglobalpulse.org.