Data Mining in Large Sets of Complex Data

Robson L. F. Cordeiro[1]
Advisor: **Caetano Traina Jr.**[1]
Co-advisor: **Christos Faloutsos**[2]

[1]Computer Science Department – University of São Paulo – São Carlos – Brazil

[2]School of Computer Science – Carnegie Mellon University – Pittsburgh – USA

robson@icmc.usp.br, caetano@icmc.usp.br, christos@cs.cmu.edu

1. Introduction

Both the amount and the complexity of the data gathered by current scientific and productive enterprises are increasing at an exponential rate, in the most diverse knowledge areas, such as biology, physics, medicine, astronomy, climate forecasting, etc. To find patterns and trends in the data is increasingly important and challenging for decision making. As a consequence, the analysis and management of *Big Data* is today a main concern of the computer science community, especially for complex data (e.g., images, graphs, audio and long texts). Finding clusters in large complex data sets is one of the main tasks in data analyses. For example, given a satellite image database with several tenths of Terabytes, how can we find regions aiming at identifying native rainforests, deforestation or reforestation? Can it be made automatically? Based on this Ph.D. work's results, the answers to both questions are a sound "yes", and the results can be obtained in just minutes. In fact, results that used to require days or weeks of hard work from human specialists can now be obtained in minutes with high precision.

Clustering complex data is a computationally expensive task, and the best existing algorithms have a super-linear complexity on both the data set cardinality (number of data elements) and dimensionality (number of attributes describing each element). Therefore, those algorithms do not scale well, precluding being efficient to process large data sets. Aimed at the analysis of *Big Data*, this Ph.D. work created new algorithms to perform clustering in moderate-to-high dimensional data involving many *billions* of elements stored in several *Terabytes* of data, such as features extracted from large sets of complex objects, but that can nonetheless be quickly executed, in just a few minutes. To achieve that performance, it was taken into consideration that high dimensional data have the clusters bounded to a few dimensions (also called axes or attributes) each, thus existing only in subspaces of the original high dimensional space, although each cluster can have correlations among dimensions distinct from those dimensions correlated in the other clusters.

In this Ph.D. research, techniques were developed to perform both "hard" and "soft" clustering (that is, assuming that each element can participate in just one or in several clusters that overlap in the data space) that can be executed by serial or by parallel processing. Moreover, their applications were shown in several practical test cases. Distinctly from most of the existing algorithms (and from all of the fastest ones), the clustering techniques developed do not require the previous definition of the number of clusters expected, rather, it is inferred from the data and returned to the user. Besides, due to the assumption that each cluster exists because of correlations existing in a subset of the space dimensions, the new techniques not only find clusters with high quality and speed, but also spot the most significant dimensions for each cluster, a benefit that the previous algorithms only achieve at the expenses of costly processing.

The methodology to develop this Ph.D. work was based on the extension of hierarchical data structures, multidimensional multi-scaling analysis of the spatial data distribution based on a convolution process using Laplacian filters [Gonzalez and Woods 2006], on the evaluation of alternative cluster entropies and on new cost functions that enable to evaluate the best strategies before executing them, allowing to perform a dynamic dataflow optimization of the parallel processing.

Our algorithm was compared with at least nine of the most efficient existing ones, and it was shown that the performance improvement is over at least one magnitude order, although *always* having its quality equivalent to the best achieved by the competing techniques. In extreme situations, it took just two *seconds* to spot clusters in real data that the best competing techniques required two *days*, with equivalent accuracy. In one of the real cases evaluated, our techniques were able to find correct tags for every image from a dataset containing many hundreds of thousand images, performing "soft" clustering (thus assigning one or more tags to each image), using as guidelines the labeling performed by a user in not more than five images for each tag (i.e., in at most 0,001% of the image set). Our experiments used up to *billions of complex objects* coming from distinct, high impact applications like breast cancer diagnosis, region detection in satellite images, assistance to climate change forecast, recommendation systems for the Web and social networks.

In summary, this Ph.D. research takes steps forward from traditional data mining (especially for clustering) by considering *large, complex datasets*. Note that, usually, current works focus in one aspect, either size or data complexity. This Ph.D. work considers both: it enables mining complex data from high impact applications; the data are large in the *Terabyte-scale*, not in Giga as usual; and very accurate results are found in just minutes. Thus, it provides a crucial and well timed contribution for allowing the creation of *real time* applications that deal with *Big Data of high complexity* in which mining on the fly can make an immeasurable difference, like cancer diagnosis or deforestation detection.

2. Basic Concepts and Related Work

Real data of dimensionality above five or so (e.g., features extracted from complex objects) tend to have many **local correlations**, as some points are commonly correlated with regard to a given set of axes, while other points are correlated regarding distinct axes [Domeniconi et al. 2007, Kriegel et al. 2009]. As a consequence, these data usually have clusters that exist only in **subspaces** of the original space (i.e., sets of orthogonal vectors formed from original axes or from subset combinations thereof) and each cluster may exist in a distinct subspace [Moise and Sander 2008, Ng et al. 2005, Moise et al. 2008, Kriegel et al. 2009]. Finding clusters in subspaces of multi-dimensional data is the goal of many works. See the Dissertation for the complete problem definition (Section 2.3) and for a recent survey of these works (Chapter 3). In summary, there are two main approaches: *bottom-up* and *top-down*. Bottom-up methods, like P3C [Moise et al. 2008] and EPCH [Ng et al. 2005], divide 1-dimensional data projections into a user-defined number of partitions and merge dense partitions to spot clusters in subspaces of higher dimensionality. Top-down methods, like LAC [Domeniconi et al. 2007] and STATPC [Moise and Sander 2008], analyze the "full dimensional" space looking for patterns that may lead to clusters. Then, the data distribution surrounding these patterns allow them to confirm the clusters and to spot their subspaces - the axes in which one cluster is denser form its subspace. Several works improve the basic ideas of the approaches top-down and bottom-up, but, unfortunately, they all remain super-linear (even with exponential complexity, for most bottom-up methods) in space or in running time.

3. Main Contributions of this Ph.D. Work

Focused at the task of clustering *Big Data* of high dimensionality, Chapter 3 of the Dissertation provides an analysis of the literature that leads to one main conclusion: Despite several qualities found in the related works, there is no previous work published in the literature and well-suited to find clusters in high dimensional data that has *any* of the following properties: (i) **linear or quasi-linear complexity** – to scale linearly or quasi-linearly in terms of memory requirement and execution time wrt increasing numbers of points and axes, and; (ii) **Terabyte-scale data analysis** – to be able to handle data of Terabyte-scale in feasible time. On the other hand, applications with Terabytes or even Petabytes of high dimensional data abound: weather monitoring systems and climate change models, where we want to record wind speed, temperature, rain, humidity, pollutants, etc; social networks like Facebook TM, with millions of nodes, and many attributes per node (gender, age, number of friends, etc); astrophysics data, such as the Sloan Digital Sky Survey, with billions of galaxies and attributes like red-shift, diameter, spectrum, etc. In fact, the analysis and management of *Big Data* is today a main concern of the computer science community, especially for high dimensional data. Thus, to overcome the two aforementioned limitations is extremely desirable nowadays.

This Ph.D. work overcomes both limitations for clustering. It was achieved with the development of three new, fast and scalable algorithms, which we describe as follows.

3.1. The Method *Halite* for Clustering Moderate-to-high Dimensional Data

Chapter 4 of the Dissertation presents *Halite*, a fast and scalable density-based clustering method that looks for clusters in subspaces of moderate-to-high dimensional data being able to analyze large complex datasets. *Halite* finds clusters based on the variation of the data density over the space in a multi-resolution way, dynamically changing the partitioning size of the analyzed regions. Multi-resolution is explored applying d-dimensional hyper-grids with cells of several side sizes over the input d-dimensional space and counting the points in each grid cell. The grid densities are stored in a quad-tree-like structure where each level represents the data as a hyper-grid in a specific resolution. A convolution process using Laplacian filters is then performed over each tree level to spot bumps in the data distribution wrt each resolution. Given a tree level, *Halite* applies a filter to find the regions in the "full dimensional" space with the largest changes in the point density. The regions found may indicate clusters that only exist in subspaces of the analyzed space. The neighborhoods of these regions are then analyzed to define if they stand out in the data in a statistical sense, thus confirming the clusters, and a compression-based analysis of the data distribution automatically spots each cluster's subspace. Finally, alternative cluster entropies are evaluated to create both "hard" and "soft" clustering results.

IMPACT: *Halite* is fast and it has linear or quasi-linear time and space complexity regarding both data size and dimensionality. Thus, *Halite* tackles the problem of **linear or quasi-linear complexity**. A theoretical study on its time and space complexity (see Section 4.3 of the Dissertation) as well as an extensive experimental evaluation (see Section 4.7 of the Dissertation) performed over synthetic and real data spanning up to 1 million objects and comparing *Halite* with seven representative works corroborate this claim. For example, *Halite* analyzed 25-dimensional data for breast cancer diagnosis (KDD Cup 2008) at least 11 times faster than the previous works, increasing their accuracy in up to 35%. Other qualities of *Halite* are: (1) *Usability:* it is deterministic, robust to noise, uses no user-defined parameter (not even the number of clusters expected), and finds clusters in subspaces formed by the original axes or by their linear combinations, allowing for space rotation; (2) *Effectiveness:* it is accurate, at least tying in clustering quality compared to top related works; and (3) *Generality:* it has both "hard" and "soft" clustering approaches.

3.2. The Method *BoW* for Clustering Big Data of Moderate-to-high Dimensionality

Given a *Terabyte-scale* dataset of moderate-to-high dimensional elements, how could one cluster them? As we discuss in Chapter 3 of the Dissertation, numerous successful, serial clustering algorithms for high dimensional data exist in literature. However, the existing algorithms, including our own method *Halite*, are impractical for datasets spanning Terabytes and Petabytes (e.g., Twitter crawl: $>$ 12 TB, Yahoo! operational data: 5 *Petabytes*[1]). Just to read a single Terabyte of data (at 5GB/min on a single modern eSATA disk) one takes more than 3 hours. For datasets that do not even fit on a single disk, parallelism is a first class option, and thus we must re-think, re-design and re-implement existing serial algorithms to allow for parallel processing. Nevertheless, good, serial clustering algorithms and strategies are still extremely valuable, because we can (and should) use them as 'plug-ins' for parallel clustering. Naturally, the best algorithm is the one that combines (a) one fast, scalable serial algorithm and (b) makes it run efficiently in parallel. This is exactly what we propose in Chapter 5 of the Dissertation.

Specifically, we explore *MapReduce* for clustering *Big Data* of moderate-to-high dimensionality. The main questions are (a) how to minimize the I/O cost, taking into account the *already existing* data partition (e.g., on disks), and (b) how to minimize the network cost among processing nodes. Either of them may be a bottleneck. To answer these questions this Ph.D. work created *BoW*, a novel, adaptive algorithm that is a hybrid between two parallel clustering strategies proposed in the Dissertation, one of the strategies minimizes I/O, while the other uses a novel *sampling-and-ignore* idea to shrink the network traffic. Both strategies have similar clustering accuracy, but the fastest one depends on the environment used. *BoW* takes the environment description as input and uses cost-based optimization to automatically choose the fastest option and proper parameters for it, prior to the real execution. Therefore, *BoW* automatically spots the bottleneck.

IMPACT: *BoW* tackles the problem of **Terabyte-scale data analysis** and, when using *Halite* as a plug-in, it also tackles the problem of **linear or quasi-linear complexity**. To corroborate this claim, Section 5.5 of the Dissertation reports experiments on real and synthetic data with *billions* of points, using more than a *thousand* cores in parallel. To the best of our knowledge, the Yahoo! web dataset (provided by Yahoo! Research) reported in the Dissertation is the largest real dataset ever reported in the clustering literature for moderate-to-high dimensional data. Spanning 0.2 TB of multi-dimensional data, *BoW* took only *8 minutes* to cluster it, using 128 cores. Further experiments used up to 1,024 cores, the highest such number in the clustering literature for moderate-to-high dimensional data. Other qualities of *BoW* are: (1) it works with most serial clustering meth-

[1] According to: Fayyad, Usama - Invited Innovation Talk - ACM SIGKDD 2007. The talk is available at "http://videolectures.net/kdd07_fayyad_dms". Day of access: June 21, 2012.

ods as a plugged-in clustering subroutine; (2) it balances I/O cost and network accesses, achieving a very good tradeoff between the two; (3) it uses no user-defined parameter; (4) it matches the serial method's clustering accuracy; and (5) it has near-linear scale-up.

3.3. The Method *QMAS* for Labeling and Summarizing Large Complex Datasets

Chapter 6 of the Dissertation provides an *additional contribution* of this Ph.D. work that is *apart from* the ones already described for clustering. It presents *QMAS*, an algorithm that uses our clustering techniques to focus on two *distinct* mining tasks – the tasks of labeling and of summarizing large complex datasets. Specifically, *QMAS* is a fast and scalable (linear) solution to two problems: (a) **low-labor labeling** – given a large set of complex objects, *very few* of which are labeled with keywords, find the most suitable labels for the remaining ones, and (b) **mining and attention routing** – in the same setting, find clusters, the top outlier objects, and the objects that best represent the main data patterns. Due to space limitations we will describe our solution to the problem of **low-labor labeling** only. It is threefold: (a) spot clusters in the input objects; (b) represent objects, clusters and known labels by distinct layers of nodes in a tri-partite graph, where object content similarities are captured by edges between object nodes and their respective cluster nodes, and the known labeled examples are represented by edges between the respective nodes of objects and keywords; and (c) for each object of interest, do random walks with restarts from this object node, thus performing a proximity query that automatically finds the best annotation keyword for the target object. In this way *QMAS* performs **low-labor labeling**.

IMPACT: *QMAS* was developed in collaboration with SAIC, a defense contractor, with (USA) national security in mind, to analyze Gigabytes of satellite images divided into millions of tiles. It allows the automatic identification of specific objects (e.g., boats, cars or buildings) in such images based on *very few* example tiles manually labeled by the user. For example, in one of the experiments performed (see Section 6.3.5 of the Dissertation), *QMAS* correctly found boats in a large set of satellite images with nearly 2.5 million tiles, using *only three examples* of tiles with boats provided by the user. *QMAS* is also being used to identify coffee crops and deforestation in satellite images from Brazil.

4. Conclusion

This Ph.D. work takes steps forward from traditional data mining by considering *large* sets of *complex data* such as images, graphs, audio and long texts. Note that, usually, current works tend to focus in one aspect, either size or data complexity. This Ph.D. research considers both: three novel methods were developed to cover clustering of large, complex datasets as well as labeling and summarization, which are even harder for this type of data; and the data are large in the *Terabyte-scale*, not in Giga as usual. The experiments consider *billions of complex data elements* (coming from different, high impact applications such as breast cancer diagnosis, region classification in satellite images, assistance to climate change forecast, recommendation systems for the Web and social networks), in which our algorithms presented very accurate results, being at least one order of magnitude faster than the best competitors. Our algorithms are open-source, available at "www.gbdi.icmc.usp.br/downloads", and they have already been downloaded nearly two dozens of times. In summary, this Ph.D. work provides a crucial and well timed contribution for allowing the development of *real time* applications that deal with *Big Data of high complexity* in which mining on the fly can make an immeasurable difference, such as cancer diagnosis or deforestation detection. For all its contributions and potential to impact on real world critical problems, for opening the door to tackle interesting future work, we believe that this work is a singular, outstanding contender for this year award.

Main publications of this Ph.D. work: The core of this work generated four main papers – one of them [Cordeiro et al. 2011a] is a regular paper at the IEEE TKDE journal (Qualis A1), one of the leading journals in databases; and the other three papers [Cordeiro et al. 2010b], [Cordeiro et al. 2011b] and [Cordeiro et al. 2010a] are published at top quality conferences, IEEE ICDE (Qualis A2), ACM SIGKDD and IEEE ICDM (Qualis A2). Additional publications are a book chapter [Traina et al. 2011], one paper in the SIAM Annual Meeting [Traina et al. 2010] and other papers in the Microsoft Research eScience Workshop, WebMedia, and SBSR cited in Section 7.5 of the Dissertation.

References

Cordeiro, R. L. F., Guo, F., Haverkamp, D. S., Horne, J. H., Hughes, E. K., Kim, G., Traina, A. J. M., Traina Jr., C., and Faloutsos, C. (2010a). QMAS: Querying, mining and summarization of multi-modal databases. In *ICDM*, pages 785–790. IEEE.

Cordeiro, R. L. F., Traina, A. J. M., Faloutsos, C., and Traina Jr., C. (2010b). Finding clusters in subspaces of very large, multi-dimensional datasets. In *ICDE*, pages 625–636. IEEE.

Cordeiro, R. L. F., Traina, A. J. M., Faloutsos, C., and Traina Jr., C. (2011a). Halite: Fast and scalable multi-resolution local-correlation clustering. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints):15 pages.

Cordeiro, R. L. F., Traina Jr., C., Traina, A. J. M., López, J., Kang, U., and Faloutsos, C. (2011b). Clustering very large multi-dimensional datasets with mapreduce. In *KDD*, pages 690–698. ACM.

Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., and Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery*, 14(1):63–97.

Gonzalez, R. C. and Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data. *ACM Transactions on Knowledge Discovery from Data*, 3(1):1–58.

Moise, G. and Sander, J. (2008). Finding non-redundant, statistically significant regions in high dimensional data. In *KDD*, pages 533–541.

Moise, G., Sander, J., and Ester, M. (2008). Robust projected clustering. *Knowledge and Information Systems*, 14(3):273–298.

Ng, E. K. K., chee Fu, A. W., and Wong, R. C.-W. (2005). Projective clustering by histograms. *TKDE*, 17(3):369–383.

Traina, A., Romani, L., Cordeiro, R. L. F., Sousa, E., Ribeiro, M., Ávila, A., Zullo Jr., J., Rodrigues Jr., J., and Traina Jr., C. (2010). How to find relevant patterns in climate data. In *SIAM Annual Meeting 2010*, page 6 pags., Pittsburgh, PA. SIAM.

Traina, A. J. M., Traina Jr., C., Cordeiro, R. L. F., Ribeiro, M. X., and Azevedo-Marques, P. M. (2011). Issues and techniques to mitigate the performance gap in content-based image retrieval systems. *Journal of Healthcare Information Systems and Informatics - IJHISI, New Tech. for Advancing Healthcare and Clinical Practices*, pages 60–83.