Data mining applications and techniques: a systematic review

Fabio T. Matsunaga[1], Jacques D. Brancher[1], Rosângela M. Busto[2]

[1] Department of Computation – Londrina State University (UEL)
Caixa Postal 6001 – 86051-990 – Londrina – PR – Brasil

[2] Center of Physical Education and Sports – Londrina State University (UEL)
Caixa Postal 6001 – 86051-990 – Londrina – PR – Brasil

ftakematsu@gmail.com, {jacques,busto}@uel.br

Abstract. *The data mining method is becoming a trend on the computer science, medicine, economy, biology, administration, environmental sciences and sport. There are several techniques and tasks related to data mining, such as clustering, classification, association rules, time series forecasting and regression model, which is being applied in several multidisciplinary problems and applications. Considering the exposed above, the aim of this paper is to present a systematic review about the data mining techniques and tasks, showing the applicability on several and multidisciplinary areas. The trend of the method, combined with other mathematical, statistical and computational methods. Also, data mining techniques and tasks are not being used in an isolated way or individually, but also as being a module or part of a project of an expert system, by combining mathematical, computational and statistical methods.*

1. Introduction

The data mining is becoming a trend on the computer science researches and other areas, such as medicine, economy, biology, administration, environmental sciences and sport. The data mining combines several techniques, such as statistic, databases and artificial intelligence. On the computer science, the technique has been used on audio/video processing, textual information extraction, data flows and web applications. The main motivation of this method using is the constant growth due to the technological advances, communication, information availability and accessibility over the Internet [Fayyad et al. 1996].

Data mining is the core of the Knowledge Discovery in Databases (KDD) process and consists in problem definition, acquisition and evaluation of data (pre-processing), unknown pattern discovering and evaluation of the resulted data for enhancement (post-processing) [Maimon and Rokach, 2010]. In other words, Data mining finds relationships and models within a large volume of data stored in a database or a technique to determine behavior in extracted information. Regardless of the algorithm, technology or technique used in essence, the result is identified patterns in a large amount of data.

The information clustering, classification and pattern recognition are one of the relevant data mining tasks [Kesavaraj and Sukumaran 2013]. The main motivation for

this is the generation of a set of rules from a known data, resulting in an organized data structure. Considering the complexity of the analysis of these structures, it is necessary to apply an artificial intelligence technique.

Considering the exposed above, the aim of this paper is to present a systematic review about state-of-art of data mining techniques and tasks, showing the applicability on several and multidisciplinary areas. Some of these tasks are classification, clustering and association rules. For this, we developed a bibliometric review presenting some trends and current problems, which are being the main targets of the researches.

This article is organized as follows: Section 2 will show the bibliometric review procedure. Section 3 will present some surveys and systematic reviews already undertaken on the research subject. In Section 4 the systematic review itself will be presented, showing the main contributions of the works. Section 5 will report about trends and the future perspective on the area. Finally, the conclusions and future works will present the conclusions and future works.

2. Bibliometric Review

To develop this review, three databases of scientific articles were consulted: IEEEXplore, Science Direct and ACM Digital Library, searching papers since 1999. Google Scholar was also used to check the citations index of the papers surveyed. The following keywords were used in the research: ”data mining”, ”data mining applications”, ”survey data mining” and ”data mining techniques” . For these keywords, more than 20,000 papers were returned on the searching process, aiding to perform quantitative statistical analysis about the scientific papers, as shown on Figure 1.

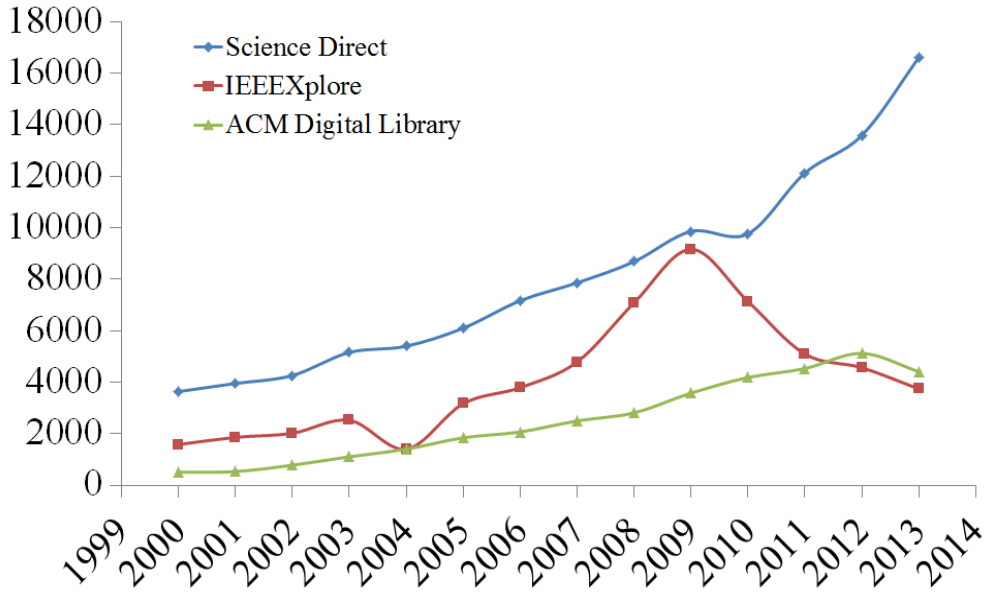

Figure 1. Quantity of articles resulted on the research on the Science Direct, IEEEXplore and ACM Digital Library database about “data mining” keyword on the last years.

In the journals and conference proceedings analyzing (Figure 1), articles were selected by reading the title and the abstract analyzing if the data mining technique was

applied in a multidisciplinary application, according to the current state-of-the-art methods. In addition, the relevance of the application results was used as a selection criterion for the systematic review development. Then, the selected articles were studied performing the review as follows: first, the problem and the research objectives, the conclusions and contributions of this study; then there was the context of practical application and inserted specific methodology or task/technique employed.

After the papers selections, an analysis of the journals with more publications on the research topic and the authors who have published on the topic ”data mining”, considering the three researches bases studies was performed. The analysis of the authors is shown in Figure 2. From that list, we selected also the latest articles of these authors.

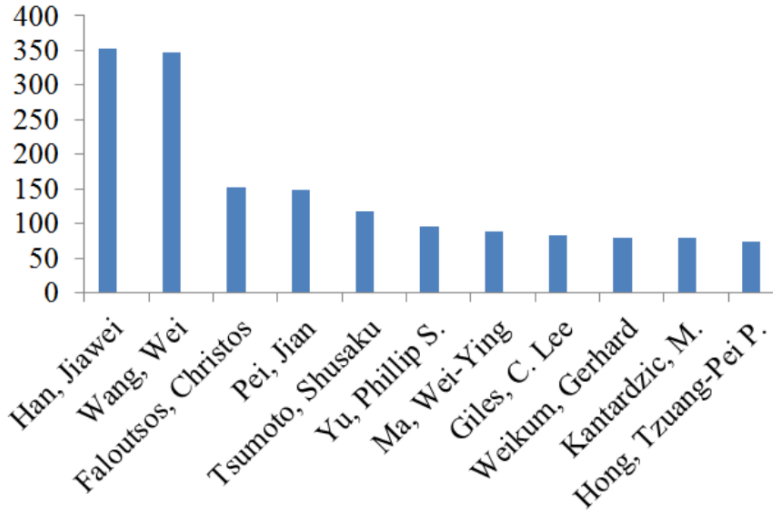

Figure 2. List of authors who published more scientific papers on Science Direct, ACM Digital Library and IEEEXplore database about “data mining” keyword.

The authors Jiawei Han and Wang Wei have more publications on the data mining area (Figure 2). The first author focused in studies involving patterns recognition techniques using data mining on network applications and the second focused in network security, big data and heterogeneous and complex data. During the 1998-2014 period, 336,757 scientific articles with “data mining” keyword were found in searching the databases IEEEXplore, Science Direct and ACM Digital Libray data.

On the Figure 1, the productions published in the Science Direct database have grown monotonically over the last 13 years (Figure 3). There was a slight decrease between 2009 and 2010, and thereafter, from the year 2011 there was a sharp increase in published articles. In contrast, the production of ACM Digital Library and IEEEXplore database were stabilized and even with some decrease in the same period prescribed for strong growth in Science Direct. Moreover, the stagnation on the IEEEXplore and ACM Digital Library occurred precisely at the same time that there was a small gap in the Science Direct in 2009.

One possible explanation for the stagnation in production and IEEEXplore, ACM Digital Library is that these databases index full papers in the area of Computer

Science, Computing and Engineering. On the other hand, the Science Direct indexes work in multiple areas of knowledge, such as Humanities, Child Health, Engineering and Earth Sciences, beyond Computer Science. This creates a chance that the techniques of data mining are being used to solve more problems in other areas, instead of improvement and study of new techniques from them.

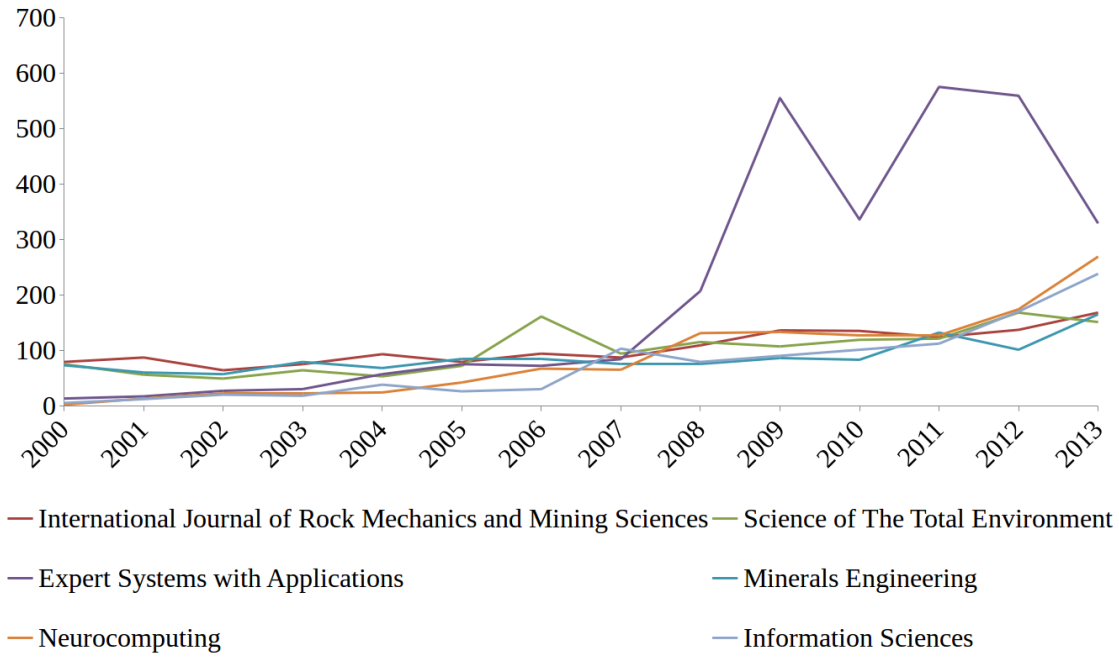

Figure 3. Number of articles with "data mining" keyword published in various journals from multiple areas indexed in Science Direct.

In Science Direct database, the journal Expert Systems with Applications (ESA) is the one that had most relevant and recent papers published that have used data mining techniques to solve applied problems. Only in ESA, there was a marked increase of articles published since 2007. This journal has considerable relevance, with an impact factor of 1.854 (with an impact factor of 2.339 in the last 5 years). Others journals with relevant results on the searching process have a considerable impact factor, as Information Sciences (3.893), Neurocomputing (2.005), Science of The Total Environment (3.163), International Journal of Rock Mechanics and Mining Sciences (1.424) and Minerals Engineering (1.714). The relevance of the topic is the fact that there is always a greater demand for the method (Figure 1), especially for the development of expert systems and intelligent (Figure 3).

After the bibliometric review, thirty eight papers were selected according to the abstract analysis. These papers were classified by the concepts applied for each paper (data mining method used) and the paper content (type of work, application environment, main methodology and if the result is positive or negative). All of these classification features were obtained by Kitchenham methodology [Kitchenham et al. 2009], which consists on make a questionnaire to classify the paper. All the content questionnaire items are shown on Table 1.

Table 1. Content data used on Kitchenham questionnaire.

	Items
Scope used	Web Mining, Text Mining, Pattern Recognition, Big Data, Machine Learning, Classification, Relational Database, Non-relational Database, Knowledge Discovery
Research type	New Method, Method Improvement, Comparison, Survey, Theoretical Research, Application
Environment applied (if application research)	Business, Biology, Chemistry, Earth Science, Environmental Science, Engineering, Medicine, Human Science
Main Methodology	Clustering, Decision Trees, Classification, Time Series Analysis, Association Rule Mining, Dependence Models, Regression Models
Result	Positive or Negative

The content questionnaire (Table 1) was one criteria used to classify the selected papers. Scopes involved the main concept of the work, different from the main methodology, which are specific computational method applied to solve the problem. Most of the research type of the papers surveyed are applications, which was the focus of this review. The result in all papers was positive, showing the feasibility of the applied methodologies on several environments.

3. Data Mining Surveys Developed

Among the surveys developed about the data mining application, Liao and Hsiao (2012) conducted a bibliometric review of scientific articles in the area from 2000 until 2011. The topics covered included types of knowledge, practical applications and architectures, which were also the basis for researches (keywords) used in the search for articles in databases. The main future trends addressed by the authors in the survey were that the data mining techniques are useful in expert systems and in developing applications. Another trend is the issue of multidisciplinarity, where areas such as psychology, cognitive and human behavior science to use the technique as a useful tool due to its large capacity for learning and adaptation to the environment.

Patil and Patil (2012) reviewed several techniques and trends about data mining method, focusing on web data, i.e., a large volume of data stored in various repositories such as data warehouses. In this case, the authors showed the importance of the current research topic, in which the data from the web is constantly growing, with a repository of information on a complex heterogeneous structure of interconnections (hyperlinks), being a very popular feature for all users. The techniques studied were: Web Content Mining (extracting information from Web documents), Web Structure Mining (information extraction structures of hyperlinks) and Web Usage Mining (information extraction from user behavior with web sites). Some future directions were established, such as enhancement algorithms, search for relevant information and discovery of new types of knowledge.

Sharma et al. (2013) surveyed various classification and machine learning techniques about data mining techniques adaptation in medical, marketing, telecommunication and health areas. Some of these techniques include support-vector machine, decision trees, Bayesian Network and *k*-nearest neighbor. The authors showed that they have different operational profiles, adapting with different performance and accuracy depending on the scenario and application requirements.

4. Data mining techniques and multidisciplinary applications

With the necessity for a new generation of computational methods and techniques for extracting useful information from a growing volume of digital data, the theory of KDD arose through a low-level mapping of the same. Thus, in the 90s, the process of data mining has emerged from the KDD theory aiming to develop a method to discover new knowledge in database [Fayyad et al. 1996].

Over the years, due to the advances of computer applications, one of the remarkable advances in theory KDD was the use of data mining on the web, from which the growth of information have become apparent [Srivastava and Cooley 2000]. This need has arisen due to the diversity of available data types (text and images), the location of the sources (hyperlinks and IP addresses) and structural organization (HTML or XML). The first practical applications that aroused interest were the electronic trades to find solutions to the challenges in the web data area.

4.1. Data mining techniques and tasks improvement

Kur-Morales and Rodríguez-Erazo (2009) proposed a new methodology for applying data mining in large databases. The approach used was the statistical reduction of the search space generating an optimized clustering model. The methodology has also been used in other areas of knowledge, such as medicine.

Another major step in the implementation of data mining is the use of database and multiple heterogeneous data [Mehenni and Moussaoui 2012], concepts widely used in e-commerce web systems. This was the main motivation to develop techniques for classification and pattern recognition, since the structure of these large databases is difficult to handle. Through this system, the authors proposed a scheme for multi-dimensional decision construction to find relations between the database trees, using a support vector regression model. However, despite the considerable results, the implemented model can be enhanced to strengthen relations between the elements, and explore other classification approaches such as artificial neural networks.

Li and Li (2013) proposed a new method of data mining task-oriented, identifying complex lithologies and reservoirs. The development of this approach is due to the great demand of parameters problems involving Earth Sciences, where the full domain involved in the data mining data-driven process are unknown, the simulations cannot produce accurate results. The use of data mining task-oriented was used with the combination of decision and support-vector machines for building a prediction model fluids trees. This research, therefore, becomes a beginning to new challenges involving data mining task-oriented.

The use of textual extraction method (text mining) to serve as input parameters for mining algorithms and classification was proposed [Williams and Gong 2014]. The aim of this research was to predict the saturation level of cost using such algorithms and procedures. The main contribution of this proposal was the issue of improving the accuracy of cost estimates, compared with other data mining models. Despite this gain, it is possible an even more accurate results if other real and additional data such as location information and complexity in learning process.

4.2. Data mining applied in Social and Human Science

Menon et al. (2005), for example, developed a system to solve conflicts during the development of complex products in a quick way. The product information is stored in a textual way in the database, from which valuable information will be extracted through data mining tasks/techniques. In the same line of Menon et al. research, Ozturk et al. (2006) explored the use of data mining with regression tree to estimate the waiting time of manufactured products. For this purpose, training data were used from pre-defined simulation models.

Still in the area of marketing, Hsu and Chen (2007) applied clustering to data mining to classify information segmentations of customers and marketing catalogs. The relevance of this work consists in the aspect of application of clustering on real data, which are mixed, variable and large. The proposed algorithm, called CAVE, served to categorize information from hierarchical distances and variance methods and entropy.

The use of data mining was still used and exploited by Alsultanny (2013) to forecast the needs of labor market. The final models resulted from processing trees decisions (for the separation of attributes), technical rules of decisions (for selective rules generation) and Naive Bayes classifiers (to create tables for the learning process, obtained from influential factors in keeping workers in jobs) were compared. Results showed that the best method was the rule-making. From this research, it was found that the use of data mining is applicable in human resources, since these manipulate large amount of information, in addition to being useful in business organizations decision making.

Hui et al. (2013) conducted an empirical research based on score applications, widely used in managing credit risks of commercial banks by data mining method. Empirical studies compared different tasks/techniques as logistic regression and other statistical analyzing. However, the research was limited only to static data and can be improved in a matter of exploring a set of large data and use of dynamic models. The advantage of the proposed system is that it performs strategic actions to improve the business organizations, analyzing the advantages and weaknesses. A future perspective addressed by the authors is that the system can be applied to other services, and application of artificial intelligence techniques for advanced analysis.

In another perspective of social science, considering the social analysis in networking, Jin et al. (2012) proposed the LikeMiner. The authors analyzed the liking social function using a mining algorithm to estimate the user interests, representativeness and influence of social media objects. The proposed work showed an effectiveness by applying the system on large scale Facebook data.

4.3. Data mining applied on Geosciences

The data mining application was also a motivation to solve geosciences problems [Bae et al. 2009]. Other studies have also continued application of the technique [Bao and He 2010], where different mining techniques have been established for handling petrophysical, geological and seismic data, to help the relationships making between information for prediction and recognition occurrence of areas with oil and minerals, even in areas with complex geological conditions reservations. Other applications, such as aid to decision making in critical systems were also assisted with the implementation of the technique [Young and Fehskens 2010].

4.4. Data Mining applied on Medicine

A work [Smith et al. 2009] applied the technique in question for the diagnoses analysis and classification, in order to obtain rules from clinical database. Statistical methods, such as bootstrap, were applied in the training data. Chen et al. (2012) developed a model of customer-oriented organizational diagnosis, called 'PARA' (primary diagnosis, advanced diagnostics, review and action), which performs identification of categories and relations between these with customers in a database of customers through data mining.

4.5. Data Mining applied in Environmental Science

Based on the Mehenni and Moussaoui work and others mentioned, the constant growth of large information motivated the development of ways of storage optimization, such as databases. Much of this information is organized with the purpose to assist in making future decisions to solve many problems, such as global disturbances and ecosystems. With the growth of information, there was also an increase in demand of input parameters, which prevented the use of conventional mathematical models and fostered the use of techniques from machine learning and data mining in these specific situations [Debeljak et al. 2014].

Other research related to high demand parameter and decision making to solve global problems was that of Perez-Palacios et al. (2014), who used the method to predict the quality of food, whose parameters were extracted from images and computer vision techniques. The multiple linear regression models were derived from the final process and allows the estimation of various parameters, such as weight, nutrient content and volumes. Thus, demonstrates its applicability in the quality control process and a motivation for its use in other cases.

Moreno-Sáez and Mora-López Saez (2014) proposed a system for simulating the distribution of solar irradiance spectra by combining the *k*-means and data mining method. These were used to group and determine the most relevant parameters and substantial solar radiation to characterize any observed spectra, from which methods regression and neural networks are used to perform the simulation based on parameters of humidity and temperatures. Thus, the main contribution of this work lies in the ability to forecast weather system through some known and available parameters.

Another problem involving the decision-making was the ease of adjustment and barriers of production plants, resulting from the incompatibility of equipment sizes and

variation in the increase in large-scale processes, which cause waste products and damages [Yang et al. 2014a]. The model generated was that of a decision tree, obtained from a method of classification and regression, obtaining combination factors and influential parameters which cause the losses. Considering the authors contributions, the same method could be applied to other types of operations, not restricted to the manufacturing area, but also the processes that constitute several steps.

A system composed of hybrid algorithms, consisting of data mining, evolution strategy and sequential quadratic programming has been developed for global optimization [Chen and Huang 2013]. This optimization consists in finding the smallest possible area of an issue, and the use of data mining in this case served to approximate the feasible regions with objective values, using the classification, association and assembly to reduce space activities that possibly contains the optimal solution.

4.6. Data mining applied in Engineering

A framework for automation systems mining, which are databases of various constructions of operations, was proposed by Xiao and Fan (2014). The framework consists of a process group for identifying patterns of energy consumption in buildings and then some rules of association were obtained to discover the relationship of energy consumption with the principal components of each group. The main contribution of this research was the use of the framework to improve the buildings performance, although it can be improved in consideration of other construction parameters to achieve better performance.

Another system was proposed using the comparison of different techniques, such as artificial neural networks, support vector machines, classification and regression trees and generalized linear regression models and multiple detector and interaction, along with the process data mining for the prediction coefficients of performance of refrigeration equipment such as R404A [Chou et al. 2014]. These coefficients were obtained from the change in the refrigerator settings, with several parameters at various stages of cooling, such as leakage of liquid and vapor. However, due to limited research in using standard refrigeration models in simple settings, the use of data mining was not always the best solution for all cases.

Another example of text classification using data mining, specifically for management of industry knowledge, was proposed by Ur-Rahman and Harding (2012). The importance of this research is due to the wide information availability in textual format, as digital documents, from which will be applied to textual data mining classification with clustering techniques and obtain useful and valuable information, which are not explicitly described. The contribution lies in the fact that the proposed methodology allows us to analyze any type of formatted text based on industrial data, despite being limited to text formatting, which creates the possibility of exploring other document formats.

4.6. Data mining applied in Networks

The traffic data over the city is one of the issues treated by pattern mining methods. This problem can be addressed in Jindal et al. (2013) work, which purpose is to develop a method to mine spatiotemporal periodic patterns in traffic data using clustering method

on road network sensors and geographical data. The main contribution of this work was the establishing of the best routes in a region.

In the same research group of Jindal et al., Khan et al. (2014) presented a tool for bugs troubleshooting due to interactive complexity in wireless networks. The distributed nature of failure scenarios motivated the sequences of events looking and discriminative sequence mining application for root cause analysis on the network. The proposed tool helped to discover the event sequences that lead to the bugs.

Yang et al., (2014b) developed a diagnostic tool with low-bandwidth radio to measure the power consumption of the host node allowing remote diagnosis. The authors solved troubles about unresponsive remote sensor nodes. The main novelty of this tool is the use of low power consumption as a side channel to diagnose sensing system failures.

5. Trends and Perspectives

A trend that can be seen from the research developed, such as Perez-Palacios et al. is that the data mining techniques and tasks are not being used in an isolated way or individually, but also as being a module or part of a project of an expert system, by combining methods and computational technologies [Chen and Huang 2013]. Furthermore, it is notable from the studies mentioned that the developed expert systems have been applied in several multidisciplinary contexts, being used as alternative to traditional tools for solving decision making.

Based on the work of Moreno-Sáez and Mora-Lopez and Perez-Palacios et al., the clustering, classification and prediction are replacing the conventional and analytical models, such as the mechanistic and analytical models. This is because of the great demand of the input parameters and information, to provide a more accurate modelling and simulation. Much of this information are real data, which in most cases do not have a well-defined pattern with no well-defined primary-key and possess various noises. These data types have a massive unstructured interconnected nature, requiring and require some complex database models, such as non-relational databases and information retrieval and natural language processing methods [Han and Wang, 2014].

Considering that most of interdisciplinary applications are composed by heterogeneous data stored in large repositories, the data mining is used in other areas due to its applicability, scientific progress and feasibility to extract and summarize useful information and knowledge, using Artificial Intelligence, Image Processing and Mathematical/Statistical methods. As mentioned by Kappoor (2014), the data mining is considered a promising trend to be applied on distributed resources, multidimensional objects, high speed data streams, noisy time series, astronomical/spatial information and World Wide Web data. These are characteristics of several multidisciplinary problems, including medical diagnosis [Kapoor, 2014], remote sensing and sensor networks [Yang et al. 2014b; Khan et al., 2014], big data information retrieval [Wang et al., 2014], decision support systems in business intelligence [Kumar et al., 2012], social networks psychological analysis [Jin et al., 2011] and other research fields.

6. Conclusions and Future Works

Considering the systematic review developed, it can be noted that the data mining techniques and tasks have been applied in many contexts, mainly to solve problems in global decision-making problems. Analyzing the bibliometric review conducted, there was a gap occurring between the years 2009 and 2010 the production of scientific articles and a continuous growth after 2010, showing the increasing of the applicability of data mining field in other areas. Nevertheless, the trend of the method, combined with other mathematical, statistical and computational methods, are being increasingly recurrent.

Moreover, in the current state-of-art data mining techniques and tasks, despite having been improved and optimized, are being applied in multidisciplinary techniques, in order to solve problems in other areas of knowledge such as medicine, human science, administration, engineering and environment science. Thus, there is no single best technique universally considerable because each of them is adapted to a particular type of problem, depending on the applied area of knowledge. It is necessary to check the requirements of the problem to be solved, how the information will be organized and pre-processed and what kind of new knowledge is desired to obtain.

As future work, we intend to use data mining techniques in real applications and scenarios, to perform the extraction of new knowledge from a set of real data. Once that data mining techniques are state-of-art in multidisciplinary applications, we intend to make its use in discovery and search for new talents from Olympic sports, which are still hidden and not explicit. This will be accomplished from the part of our research, for web mining about sports results automatically retrieved and extracted from the web for further social and psychological analysis of the athletes.

Acknowledgements

This work has received funding from the CNPq project number 487430/2013-1. Fabio Takeshi Matsunaga is supported by CNPq grant, process number 381241/2014-9.

References

Alsultanny, Y. a. (2013). Labor Market Forecasting by Using Data Mining. *Procedia Computer Science*, 18:1700–1709.

Bae, D.-H., Baek, J.-H., Oh, H.-K., Song, J.-W., and Kim, S.-W. (2009). SD-Miner: A spatial data mining system. In 2009 *IEEE International Conference on Network,* pages 803-807.

Infrastructure and Digital Content, pages 803–807. Ieee. Bao, F. and He, X. (2010). Applying data mining to the geosciences data. In *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering*, pages 290–293.

Chen, C.-K., Shie, A.-J., and Yu, C.-H. (2012). A customer-oriented organizational diagnostic model based on data mining of customer-complaint databases. *Expert Systems with Applications*, 39(1):786–792.

Chen, T. and Huang, J. (2013). Application of data mining in a global optimization algorithm. *Advances in Engineering Software*, 66:24–33.

Chou, J.-S., Hsu, Y.-C., and Lin, L.-T. (2014). Smart meter monitoring and data mining techniques for predicting refrigeration system performance. *Expert Systems with Applications*, 41(5):2144–2156.

Debeljak, M., Poljanec, A., and Zenko, B. (2014). Modelling forest growing stock from inventory data: A data mining approach. *Ecological Indicators*, 41:30–39.

Fayyad, U., Piatetsky-shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37–54.

Han, J. and Wang, C. (2014). Mining latent entity structures from massive unstructured and interconected data. In *Proceedings of the 2014 ACM SIGMOD international conference on management of data*, pages 1409–1410.

Hsu, C.-C. and Chen, Y.-C. (2007). Mining of mixed data with application to catalog marketing. *Expert Systems with Applications*, 32(1):12–23.

Hui, L., Li, S., and Zongfang, Z. (2013). The Model and Empirical Research of Application Scoring based on Data Mining Methods. *Procedia Computer Science*, 17:911– 918.

Jin, X., Wang, C., Luo, J., Yu, X. and Han, J. (2011). LikeMiner: a system for mining the power of 'like' in social media networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 753–756.

Jindal, T., Giridhar, P., Tang, L., Li, J. and Han, J. (2013). Spatiotemporal periodical pattern mining in traffic data. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing,* pages 1–8.

Kapoor, A. (2014). Data Mining: Past, Present and Future Scenario. *International Journal of Emerging Trends & Technology in Computer Science*, 3(1):95–99.

Kesavaraj, G. and Sukumaran, S. (2013). A study on classification techniques in data mining. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–7.

Khan, M. M. H., Le, K. H., Ahmadi, H., Abdelzaher, T. F. and Han, J. (2014). Troubleshoooting interactive complexity bugs in wireless sensor networks using data mining techniques. *ACM Transactions on Sensor Networks*, 10(2):1-35.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology* 51 (1):7–15.

Kuri-Morales, A. and Rodríguez-Erazo, F. (2009). A search space reduction methodology for data mining in large databases. *Engineering Applications of Artificial Intelligence*, 22(1):57–65.

Kumar, P., Nitin, Chauhan, D. S. and Sehgal, V. K. (2012). Selection of evolutionary approach based on hybrid data mining algorithms afor decision support system and business intelligene. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*. pages 1041–1046.

Li, X. and Li, H. (2013). A new method of identification of complex lithologies and reservoirs: task-driven data mining. *Journal of Petroleum Science and Engineering*, 109:241–249.

Liao, S.-H., Chu, P.-H., and Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303–11311.

Maimon, O. and Rokach, L. (2010). Introduction to knowledge discovery and data mining. In *Data Mining and Knowledge Discovery Handbook*, pages 1-15, Springer US.

Mehenni, T. and Moussaoui, A. (2012). Data mining from multiple heterogeneous relational databases using decision tree classification. *Pattern Recognition Letters*, 33(13):1768–1775.

Menon, R., Tong, L. H., and Sathiyakeerthi, S. (2005). Analyzing textual databases using data mining to enable fast product development processes. *Reliability Engineering & System Safety*, 88(2):171–180.

Moreno-Sáez, R. and Mora-López, L. (2014). Modelling the distribution of solar spectral irradiance using data mining techniques. *Environmental Modelling & Software*, 53:163–172.

Ozturk, A., Kayaligil, S., and Ozdemirel, N. E. (2006). Manufacturing lead time estimation using data mining. *European Journal of Operational Research*, 173(2):683–700.

Patil, U. M. and Patil, J. B. (2012). Web data mining trends and techniques. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics - ICACCI '12*, page 961, New York, New York, USA. ACM Press.

Pérez-Palacios, T., Caballero, D., Caro, A., Rodríguez, P. G., and Antequera, T. (2014). Applying data mining and Computer Vision Techniques to MRI to estimate quality traits in Iberian hams. *Journal of Food Engineering*, 131:82–88.

PhridviRaj, M. and GuruRao, C. (2014). Data Mining – Past, Present and Future – A Typical Survey on Data Streams. *Procedia Technology*, 12:255–263.

Sharma, S., Agrawal, J., Agarwal, S., Sharma, S. (2013). Machine Learning Techniques for Data Mining: A Survey. In *IEEE International Conference on Computational Intelligence and Computing Research,* pages 1-6.

Smith, M., Wang, X., and Rangayyan, R. (2009). Evaluation of the sensitivity of a medical data-mining application to the number of elements in small databases. *Biomedical Signal Processing and Control*, 4(3):262–268.

Srivastava, J. and Cooley, R. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations*, 1(2):12–23.

Ur-Rahman, N. and Harding, J. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, 39:4729–4739.

Wang, L., Lin, J., Metzler, D. and Han, J. (2014). Learning to efficiently rank on big data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 209–210.

Williams, T. P. and Gong, J. (2014). Automation in Construction Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction*, 43:23–29.

Xiao, F. and Fan, C. (2014). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, 75:109–118.

Yang, Y., Farid, S. S., and Thornhill, N. F. (2014). Data mining for rapid prediction of fa-cility fit and debottlenecking of biomanufacturing facilities. *Journal of Biotechnology*. 179:17-25.

Yang, Y., Su, L., Khan, M., Lemay, M., Abdelzaher, T., Han, J. (2014). Power-Based Diagnosis of Node Silence in Remote High-End Sensing Systems. *ACM Transactions on Sensor Networks,* 11(2): 1–33.

Young, T. and Fehskens, M. (2010). Utilizing data mining to influence maintenance actions. In *AUTOTESTCON*, number 10, pages 1–5.