What is the importance of data mining for logistics and supply chain management? A bibliometric review from 2000 to 2014

Roberto Fray da Silva (roberto.fray.silva@gmail.com)
Carlos Eduardo Cugnasca
School of Engineering of the University of São Paulo, Brazil
Isabel Praça
Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, Portugal

Abstract
Data mining extracts knowledge from large volumes of data. Bibliometrics was used to identify journals, papers, authors and events that applied it to logistics and to supply chain management. 255 documents were found and analyzed. Three clusters were identified: theory development, market-oriented solutions, and theory application. The relevance of the technique was discussed.

Keywords: bibliometrics, data mining, logistics

Introduction

Logistics is the field that studies and develops solutions to improve products, information and money flows. It is divided into three main areas: facility location, product transportation and delivery, and warehousing. During the 1990s, the concept of supply chain management (SCM) was developed to consider the interactions between the different companies in the process of satisfying consumer demands (BALLOU, 2006; CHOPRA, MEINDL, 2010).

According to Chopra and Meindl (2010) and Lee, Padmanabhan and Whang (2004), some of the important problems that occur in a supply chain, such as the bullwhip or Forrester effect, are due to the lack or asymmetry of information among its links. Although these concerns are highly relevant, it is also worth considering how this information is going to be collected, in terms of technology, and how the huge number of data generated will be processed and transformed into knowledge for the companies involved in the supply chain.

There is a need to develop computational theories and tools that will allow the transformation of all the data gathered by companies into knowledge that will be useful to their decision-making processes. Fayyad, Piatetsky-Shapiro and Smyth (1996) define data mining as the application of specific algorithms to extract patterns from data. This process mainly involves data clustering, classification and regression, and the final goal is to extract relations from preprocessed data.

As it allows extracting information and knowledge from huge volumes of raw data, data mining can assist decision-making and strategic planning in areas such as logistics and SCM. In the case of SCM, the knowledge generated could benefit the whole supply chain, not only one of its links, improving its overall effectiveness.

For this reason, the main objective of this paper is to review the scientific literature related to data mining, logistics and SCM areas to identify the main clusters related to the use of data mining techniques in these areas. A secondary goal is to identify the main journals,

conferences, authors and papers related to these subjects. The results will provide a framework for further research, in the format of a cluster analysis.

Data mining

Data mining and knowledge discovery use methods, algorithms, and specific techniques to extract useful information from data, and can be divided into five steps: data selection, preprocessing, transformation, data mining, and interpretation. These steps are illustrated in Figure 1 (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996, PENG et al, 2008).

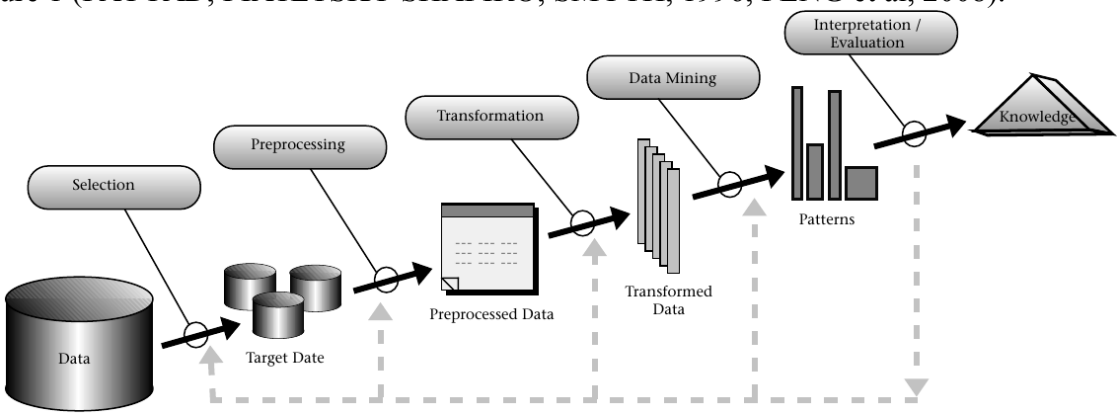

Figure 1. An overview of the steps that compose the knowledge discovery process. Source: Fayyad, Piatetsky-Shapiro and Smyth, 1996.

Note that data mining is a process that occurs with data that is already transformed, with non-relevant information already excluded from the original dataset. The patterns identified by this process will then be interpreted and evaluated based on the specific context, and knowledge that may assist decision making will be obtained.

According to Peng et al (2008), the main techniques used in data mining to extract patterns from data are: classification rules or trees, regression, and clustering. This paper uses clustering techniques and a specific algorithm to extract information from preprocessed data (the relevant papers selected from a scientific database). The raw data, in the case of this paper, is related to the papers found in the database after a simple keyword search, and the preprocessing is related to eliminating non-relevant papers.

Thuraisingham (2000), on the other hand, classifies the main techniques used in data mining into six categories: classification, association, clustering, prediction, and estimation and deviation analysis. Classification, the simplest of them, groups items based on a specific attribute. Association deals with discovering relationships between items. Clustering is a method for discovering groups that contain different items with similar characteristics. Prediction and estimation deal with trend analysis, seeking to forecast values and deduce other attributes, respectively. Finally, deviation analysis is related to developing comparisons between the data and a standard to identify anomalies.

Because data clustering is an important part of this research, it deserves further definition. Data clustering, according to Jain, Murty and Flynn (1999) can be described as an unsupervised classification of preprocessed data into groups, called clusters. These groups share commonalities, which are not shared with other clusters, and specific algorithms are used in the clustering method. In the case of this paper, an algorithm that is widely accepted to identify patterns in scientific papers is used, and the researchers are more interested in the results of the process (the clusters identified and their components) than in the clustering method itself (the steps followed by the algorithm).

Data mining techniques are important to extract information from datasets that contain enormous amounts of data. Tseng et al (2006), for example, use a rough-set algorithm together with the support vector machine method to assist the selection of suppliers for a company. One of the reasons to adopt data mining techniques is related to the number of suppliers and criteria to be evaluated for each, a recurring theme in logistics and SCM.

Logistics and supply chain management

As stated before, logistics can be defined as a multidisciplinary field that studies and develops solutions to improve products, information and money flows, both regarding product and service industries. It is divided into three main areas: facility location, product transportation and delivery, and warehousing, and seeks mainly to optimize the current situation. (BALLOU, 2006).

Data mining can have multiple uses for helping to address logistics problems: in facility location, it can help to identify the optimal location, which may minimize costs or maximize profit, or how to best allocate production in different facilities over time; in the case of transportation, it can assist with selecting the transporters and logistics service providers, with evaluating their performance, and with scheduling vehicles during product distribution; and in warehousing, it can help to monitor products in different locations, and to identify their consumption patterns, among others.

The use of data mining in a supply chain can bring even more benefits when compared to logistics, such as allowing the identification of trends that influence all its links, which supply chain design better fits specific consumer target groups, how to reduce the effects of the bullwhip effect, among others. Even though these techniques may directly affect the companies' results, there is a lack of scientific papers that aim to identifying the state of the art of the data mining application in logistics and SCM.

Methodology

This paper can be described as an exploratory and descriptive study because it seeks to develop a general structure to be analyzed in depth in future studies and to better describe a subject. The methodology adopted was the bibliometric analysis or bibliometric review, due to its advantages to reach the objective proposed in comparison with the main alternative, a standard literature review. Several papers used this methodology, obtaining interesting results, such as: Wormell (2000), Qiu and Chen (2009), Elm et al. (2009), and Silva et al. (2014).

The methodology used can be divided into five steps that will be described in depth in the following paragraphs, to allow further replication of the research.

The first step was the identification of the main keywords that are representative of the data mining and logistics areas and supply chain management. Experts from the Agricultural Automation Laboratory (LAA, part of the University of São Paulo) and the Knowledge Engineering and Decision-Support Research Center (GECAD, part of the Porto Polytechnic Institute) were consulted at this step.

The second step was a search into the Web of Science database, between 2000 and 2014, using the keywords defined at the first step: clustering, classification, regression, deviation analysis, decision trees, artificial neural networks, genetic algorithms, clustering algorithms, fuzzy logic, data mining, decision support systems and knowledge discovery, together with logistics or supply chain management. This search resulted in 392,909 documents.

The third step was related to filtering the results found to improve their accuracy. First, only articles, meetings and reviews were selected, resulting in 300,290 documents.

Then, a language filter was applied, selecting only papers in English or Portuguese, and narrowing the results to the "Science and Technology" and "Social Sciences" domains of the Web of Science Core Collection database, resulting in 8,328 documents. The last filter applied was related to the areas of knowledge, selecting only papers that belonged to the following areas: computer science, engineering, business economics, operations research management science, other topics in science technology, transportation, automation control systems, agriculture, and other topics in social sciences. This resulted in 2,991 documents.

The results were then ranked from the highest to the lowest number of citations, and divided in six PDF files, to facilitate their analysis. These were then analyzed and selected based on the fit of their title and abstract with the research objectives. This resulted in 500 documents. To further improve the quality of the research, only papers with one or more citations were selected, resulting in 255 documents.

In the fifth step, the documents were analyzed using the VosViewer software (Eck and Waltman, 2009), which uses an algorithm to analyze the similarities between words in a paper title and abstract and generate a cluster map. A thesaurus with 47 entries was elaborated to avoid double counting of terms, and incorrect counting of different words related to the same terms. The clusters identified were used to analyze the papers, along with the following criteria: relevance, use of data mining, and number of citations. An analysis of the most cited papers, journals, and events was performed using Microsoft Excel.

Results and discussion

This section of the paper contains the results observed by the researchers. It is divided into sample description, in which the main data that characterize the 255 relevant papers is summarized, and cluster analysis, which contains the clusters identified, the main keywords, and examples of relevant papers.

Sample description

The sample analyzed herein was formed by the 255 most relevant papers identified. A description of this sample will be provided on the following paragraphs.

Figure 2 contains the number of papers published per year, from January 2000 until December 2014. The most relevant years, in number of publications, were observed to be the ones between 2008 and 2010. This period accounted for 44% of the sample, or a total of 111 papers. The year of 2009 alone corresponds to 16% of the sample.

It is also possible to infer that data mining application to the logistics and SCM areas became more important after 2007, because the interval between 2007 and 2014 contains 73% of the papers of the whole sample. Several reasons can explain this observation: newer data mining techniques, easier to use software, higher technology adoption by companies in different sectors, and the increasing importance of SCM, among others.

Tables 1 and 2 contain the most important journals in the sample analyzed in terms of the number of citations and number of papers published. The main journals considering both aspects, which can be inferred as being the most important for this research, are: OR Spectrum, European Journal of Operational Research, International Journal of Logistics Management, International Journal of Physical Distribution & Logistics Management, Transportation Research Record and Transport.

Table 3 lists the 10 main authors in the sample, considering the number of citations. The calculation followed the method used by Silva et al (2014), in which the authors' names are standardized, and then the number of citations per author is added using a pivot table in Microsoft Excel. Then, the table is ordered from the most cited to least cited authors. Note

that the first three authors have considerably more citations than the other 7 in the list, totaling 69% of the citations of the 10 most cited authors.

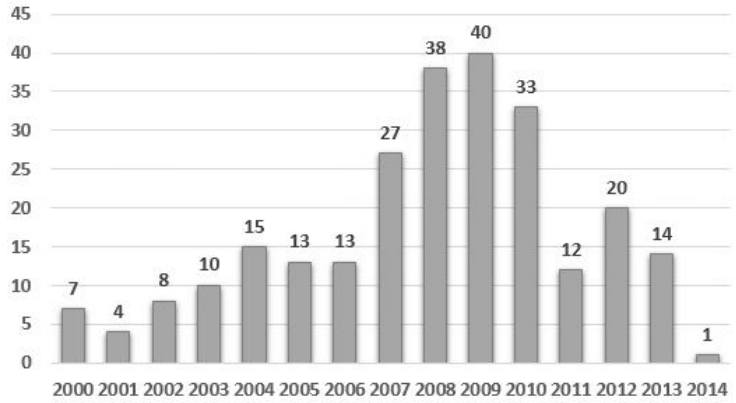

Figure 2 - Number of papers per year in the sample analyzed, from 2000 until 2014

Table 1 - Main journals in the sample analyzed in numbers of citations

Number	Magazine	Number of papers
1	OR Spectrum	813
2	European Journal of Operational Research	217
3	International Journal of Logistics Management	66
4	Journal of Computer Information Systems	49
5	International Journal of Physical Distribution & Logistics Management	46
6	Transportation Journal	41
7	Transportation Research Record	37
8	Transport	27
9	Journal of Business Logistics	26
10	Journal of Applied Research and Technology	18

Table 2 - Main journals in the sample analyzed in numbers of papers published

Number	Magazine	Number of papers
1	OR Spectrum	11
2	Transportation Research Record	6
3	African Journal of Business Management	5
4	European Journal of Operational Research	5
5	Transport	5
6	International Journal of Logistics Management	3
7	International Journal of Physical Distribuition & Logistics Management	3
8	International Journal of Transport Economics	3
9	Promet - Traffic & Transportation	3

Table 4 contains the 10 most cited papers in the sample analyzed. Most of them are verified to be related to physical distribution, logistics management and operations research. They focus mainly on physical products production, warehousing and distribution. None of the scientific journals focus on services.

Table 5 contains the most important events identified in the sample analyzed. It allows observing that they focus mainly on the following areas: city logistics, simulation, engineering and management.

Table 3 - Main authors in the sample analyzed in numbers of citations as first authors

Number	Author	Number of citations
1	Steenken, D.	341
2	Stahlbock, R.	246
3	Timpe, C. H.	103
4	van der Vorst, J. G. A. J.	71
5	Fabbe-Costes, N.	47
6	Gen, M.	45
7	Stadtler, H.	44
8	Knemeyer, A. M.	41
9	Garcia-Flores, R.	33
10	Hartmann, S.	33

Table 4 - Main papers and total number of citations in the sample analyzed

Authors	Title	Magazine	Number of citations
Steenken, D.; Voss, S.; Stahlbock, R.	Container terminal operation and operations research - a classification and literature review	OR Spectrum	341
Stahlbock, R.; Voss, S.	Operations research at container terminals: a literature update	OR Spectrum	246
Timpe, C.H.; Kallrath, J.	Optimal planning in large multi-site production networks	European Journal of Operational Research	103
van der Vorst, J.G.A.J.; Beulens, A.J.M.; van Beek, P.	Modelling and simulating multi-echelon food systems	European Journal of Operational Research	68
Fabbe-Costes, N.; Jahre, M.	Supply chain integration and performance: a review of the evidence	International Journal of Logistics Management	47
Gen, M.; Altiparmak, F.; Lin, L.	A genetic algorithm for two-stage transportation problem using priority-based encoding	OR Spectrum	45
Stadtler, H.	A framework for collaborative planning and state-of-the-art	OR Spectrum	44
Knemeyer, A.M.; Murphy, P.R.	Exploring the potential impact of relationship characteristics and customer attributes on the outcomes of third-party logistics arrangements	Transportation Journal	41
Hartmann, S.	A general framework for scheduling equipment and manpower at container terminals	OR Spectrum	33
Garcia-Flores, R.; Wang, X.Z.	A multi-agent system for chemical supply chain simulation and management support	OR Spectrum	33

Table 5 - Main events in the sample analyzed in numbers of papers

Number	Event	Number of papers
1	7th International Conference on City Logistics	7
2	6th International Conference on City Logistics	5
3	2003 Winter Simulation Conference	4
4	2007 IEEE International Conference on Industrial Engineering and Engineering Management	3
5	2008 IEEE International Conference on Service Operations and Logistics, and Informatics	3

Cluster analysis

This section contains an analysis of the three clusters identified and suggestions for further research.

Figure 3 shows the three clusters identified, with their most important keywords. This image was generated using the VosViewer software, using its default configuration and algorithm, and the analysis was based on the text in the abstract section of the 255 papers analyzed. The clusters are observed to be quite simple in terms of number of keywords, if compared to similar reviews conducted by other researchers. Nevertheless, there is a clear distinction between the clusters in terms of content.

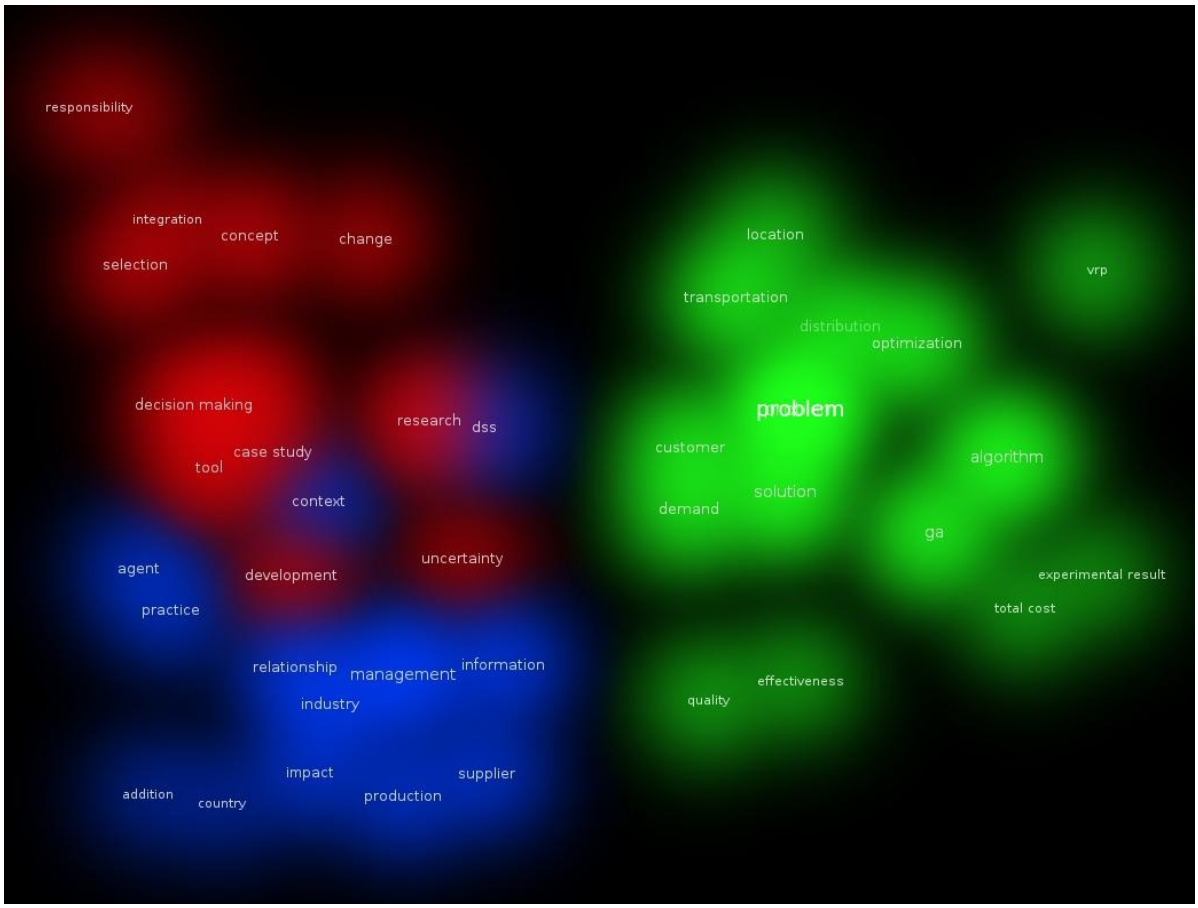

Figure 3 - Clusters generated by the analysis of the sample with the Vosviewer software

The first cluster identified, the green colored one in Figure 3, was named "Theory Development" because it contains terms and concepts related to theory development, modeling and experiments in the application of data mining techniques. This is the most important cluster, with most of the cited papers in the sample analyzed. From the 20 most cited papers in the sample, 12 papers, or 60%, belong to this cluster. It describes what techniques can be used for specific cases, such as container terminals operation, container transportation, vehicle routing, and optimal site planning. Among the main techniques used by the papers in this cluster, these can be cited: genetic algorithms, heuristics, and optimization models applied to the three main logistics problems. The main keywords in this cluster are: problem, location, transportation, distribution, optimization, vrp, customer, demand, solution and algorithm.

Steenken, Voß and Stahlbock (2004) conducted an in-depth literature review on container terminals optimization. The logistics operations involved with container handling generated a considerable number of data that has to be considered in the decision making process, such as: how to handle the empty containers, where to store the containers in the terminal, how to arrange them on the ships, what the best routes for the ships are, how to deal with containers that have special needs (related to temperature maintenance or need to isolate

from other containers), among others. Data mining algorithms can help the decision maker to choose actions related to all these problems by extracting trends from the data and suggesting options that will optimize the systems.

Gen, Altiparmak and Lin (2006) develop a genetic algorithm to optimize a two-stage transportation problem that includes three different problems: how many distribution centers to open, how the shipments between the plants to the distribution centers should occur, and how these should be made from the distribution centers to the customers. Timpe and Kallrath (2000) develop a linear mixed integer model that also considers the setup costs in multi-purpose plants, facilities that have equipment that can be operated in different ways. This adds complexity to the original problem, considering and even higher number of data.

Stadtler (2009) conducted a literature review on logistics collaborative planning, concluding that the needs for the different supply chain links need to be aligned so that the supply chain can be optimized. Collaborative planning, according to Lee, Padmanabhan and Whang (2004) and Stadtler (2009) may help reducing the information asymmetry in the supply chain, improving its results. However, this demands analyzing a considerable number of data, on a regular basis; data mining techniques, coupled with mathematical modeling, can significantly improve the results and the response time of these analyses.

The second cluster, named "Market-oriented Solutions" (the red colored cluster in Figure 3) is related to researches that focus on developing decision making models and decision support systems, applying real market constraints to the theories developed at the first cluster. The main methodology used by the papers in this cluster is the case study methodology, making them more specific in terms of scope. The main goal of these papers is to bring theory into practice, applying models and frameworks to real cases. The main keywords in this cluster are: decision-making, case study, tool, research, concept, change, integration, uncertainty, and selection.

Almeida (2001) studies the repair contract selection and spares provisioning problems in a factory, a typical operation problem that involves not only lots of data, but also different objectives. He then develops a multicriteria decision-making model to assist decision makers to evaluate options related to both the costs and risks dimensions.

Everett (2001) conducts a research with characteristics from both a literature review and a case study, to identify how algorithms and simulation models can help improve iron ore quality, related to product composition, in the production stage. This stage is characterized by four scheduling operations that may affect product composition. The author studied possible solutions for improving quality in two companies for several years, and identified the main bottlenecks for optimizing these operations. Berning et al (2003) consider a similar problem in the chemical industry, but they prefer to use a genetic algorithm to search for the optimal solution. Their mathematical model also considers alternative production pathways and software was developed to apply it to the supply chain.

Also related to the chemicals supply chain, a research developed by García-Flores and Wang (2002) elaborates a multi-agent system to simulate and to support management decisions. A tool that can be used on the internet was designed, and agents can communicate with each other, lowering problems of information asymmetry.

The last cluster, the blue one in Figure 3, was named "Theory Applications", because it focuses on identifying the main factors that affect the companies' results, including analysis of decision-making models, decision support systems, and stakeholder management. Unlike the other clusters, mathematical models are not used as the main pillars for problem solving. The main keywords in this cluster are: management, information, industry, relationship, practice, agent, impact, dss, production, and supplier.

Pan and Jang (2008) use the technology-organization-environment framework to evaluate the factors that determined adoption of the enterprise resource planning (ERP)

software by Taiwan communications industry, by interviewing 99 firms. Regression analysis was used to process the data, and the resulting model was shown to explain 79% of the decisions made. The main factors identified were: technology readiness, production and operations improvement, size of the company, and perceived adoption barriers.

Zhou et al (2008) used the data envelopment analysis (DEA) methodology to analyze the operational efficiency of ten important Chinese third-party logistics providers. Regression analysis was used to identify the main factors that influenced the performance of those companies. The authors observed that a decline in efficiency in the period studied was related to three main factors: the SARS outbreak, the difficulty in adapting to the market-based economy, and the type of services provided. They identified that sales opportunity and the level of technical expertise are relevant factors for these companies' efficiency measurements.

Limitations and future research

The main limitations observed in this research can be divided into two categories: methodology-related and content-related. The first category contains limitations that are common to the use of this specific methodology: the keywords selected and the period chosen for analysis. The second category limitations are related to the criteria chosen to evaluate the relevant papers.

Overall, the bibliometric review allowed gathering insights that would hardly be observed by conducting a standard literature review, because it allows the analysis of a considerably larger number of papers. The researchers believe that, because the research group had experts in both areas analyzed, the limitations were minimized.

The identified clusters are currently being analyzed in more depth, along with the papers that compose them. This will allow identifying the main methodologies and data mining techniques used for solving each of the three logistics problems, how data mining was used during the evolution of the logistics and SCM concepts, and what the main gaps are in terms of applying data mining to the logistics and SCM areas.

Conclusions

In this paper, the bibliometric review method was used to identify the main clusters and keywords related to the use of data mining techniques in logistics and SCM. The main journals, conferences, authors and papers were identified, providing the basis for more in-depth studies on the use of data mining in these areas. The research identified 255 relevant papers that deal with aspects of production, transportation, supplier selection, warehousing, facility location, vehicle routing, among others.

The cluster related to theory development was the most important in terms of number of papers, and the cluster related to theory application was composed mainly of case studies. Papers on all three levels of logistics problems were found: strategic, tactical and operational.

The main limitations of this research are the keywords selected, the period chosen for analysis and the criteria chosen to evaluate the relevant papers. Further research already being conducted relates to the analysis of each cluster in depth, identifying methodologies and the data mining techniques used to solve each of the logistics problems, along with identifying the main gaps in the application of these techniques.

Bibliography

Almeida, A. T. 2001. Multicriteria decision making on maintenance: spares and contracts planning. **European Journal of Operational Research**, 129(2), 235-241.

Ballou, R. H. 2006. **Gerenciamento da Cadeia de Suprimentos**. Editora Bookman, Porto Alegre – RS.

Berning, G., Brandenburg, M., Gürsoy, K., Mehta, V., Tölle, F. J. 2002. An integrated system solution for supply chain optimization in the chemical process industry. **OR spectrum**, 24(4): 371-401.

Chopra, S., Meindl, P. 2010. **Supply Chain Management: Strategy, Planning and Operation**. Fourth Edition, Pearson Education, New Jersey, USA.

Eck, N. J., Waltman, L. 2009. Vosviewer: a computer program for bibliometric mapping. **In:** Proceedings of the 12th International Conference on Scientometrics and Infometrics: 886-897.

Elm, E., S. Wanlder, P. Juni. 2009. The role of correspondence secions in post-publication peer review: a bibliometric study of general and internal medicine journals. **Scientometrics** 81(3): 747-755.

Everett, J. E. 2001. Iron ore production scheduling to improve product quality. **European Journal of Operational Research**, 129(2), 355-361.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. 1996. From data mining to knowledge discovery in databases. **AI Magazine**, 17(3), 37.

García-Flores, R., Wang, X. Z. 2002. A multi-agent system for chemical supply chain simulation and management support. **OR Spectrum**, 24(3), 343-370.

Gen, M., Altiparmak, F., Lin, L. 2006. A genetic algorithm for two-stage transportation problem using priority-based encoding. **OR Spectrum**, 28(3), 337-354.

Jain, A. K., Murty, M. N., Flynn, P. J. 1999. Data clustering: a review. **ACM Computing Surveys (CSUR)**, 31(3), 264-323.

Lee, H. L., Padmanabhan, V., Whang, S. 2004. Information distortion in a supply chain: the bullwhip effect. **Management Science**, 50(12), 1875-1886.

Pan, M. I. N. G. J. U., Jang, W. 2008. Determinants of the adoption of enterprise resource planning within the technology-organization-environment framework: Taiwan's communications. **Journal of Computer Information Systems**, 48(3).

Peng, Y., Kou, G., Shi, Y., Chen, Z. 2008. A descriptive framework for the field of data mining and knowledge discovery. **International Journal of Information Technology & Decision Making**, 7(04), 639-682.

Qiu, H., Chen Y. 2009. Bibliometric analysis of biological invasions research during the period of 1991 to 2007. **Scientometrics**, 81(3): 601-610.

Silva, R. F., Amato, J., Yoshizaki, H. T. Y., Cugnasca, C. E. (2014). The state of the art in cleaner production: a bibliometric analysis from 2000 until 2013. **In:** Proceedings of the 2014 POMS Conference, Production and Operations Management Society, Atlanta, USA, 1-10.

Stadtler, H. 2009. A framework for collaborative planning and state-of-the-art. **OR Spectrum**, 31(1), 5-30.

Steenken, D., Voß, S., Stahlbock, R. 2004. Container terminal operation and operations research-a classification and literature review. **OR Spectrum**, 26(1), 3-49.

Thuraisingham, B. 2000. A primer for understanding and applying data mining. **IT Professional**, 2(1): 28-31.

Timpe, C. H., Kallrath, J. 2000. Optimal planning in large multi-site production networks. **European Journal of Operational Research**, 126(2), 422-435.

Tseng, T. L., Huang, C. C., Jiang, F., Ho, J. C. 2006. Applying a hybrid data-mining approach to prediction problems: a case of preferred suppliers prediction. **International Journal of Production Research**, 44(14), 2935-2954.

Wormell, I. 2000. Bibliometric analysis of the welfare topic. **Scientometrics**, 48 (2): 203-236.

Zhou, G., Min, H., Xu, C., Cao, Z. 2008. Evaluating the comparative efficiency of Chinese third-party logistics providers using data envelopment analysis. **International Journal of physical distribution & logistics management**, 38(4), 262-279.