v. 2, n. 7, 2022

Journal of
Engineering Research

KNOWLEDGE EXTRACTION FROM SCIENTIFIC ARTICLES: A PROPOSAL FOR INITIAL RESEARCH METHODOLOGY ON TEXT MINING

Edgar Marcos Ancioto Junior
Pontifícia Universidade Católica de Goiás

Maria José Pereira Dantas
Pontifícia Universidade Católica de Goiás

Abstract: The need to extract textual information encourages the development of automated tools for document reading, pattern recognition and knowledge extraction. The large amount of scientific content available on the internet makes it difficult to search for the most relevant and up-to-date information. In this work, a research methodology is performed to support the process of information retrieval in a scientific environment with the initial theme of Text Mining. Steps were defined for searching, visualizing and filtering data, through computational tools, with the objective of finding relevant scientific material with assertiveness. The method resulted in a list of 49 published articles with an average impact factor of 2.193, material that was compiled and used for initial mapping of the topic. It is concluded that the initial research methodology proposal provides results with a reduced time, with relevance to the material and that the use of general-purpose computational tools, even with a lot of manual intervention, helps to define the initial state of a scientific research. Text mining can make the extraction of knowledge from scientific research already produced more productive, promoting the identification of the state of the art and research gaps.
Keywords: Bibliometrics; Web of Science Base; Automatic extraction; Vosviewer software.

INTRODUCTION

The continuous growth of information available on the Internet along with the trend of digitization of the modern world has led to a huge increase in existing data (WONG, 2012), where the user is faced with difficulties in obtaining relevant results in their research. The ability to generate this data evolves every day, which requires the emergence of methods to obtain, organize and facilitate access to this volume of data.

Techniques and methods help in the information retrieval process, use pattern matching and keyword combinations (WONG, 2012), but it is not yet developed enough to provide the existing concepts about the relationships between data (ABULAISH; ANWAR, 2012). This way, the available textual data is associated with the access challenge due to its unstructured nature.

Among these textual data, scientific publications in the form of technical reports, journal and conference articles, dissertations and theses stand out, where part of this literature is published electronically. The large amount of scientific content available on the Internet makes it difficult to search for the most relevant and up-to-date information (JOORABCHI, ARASH; MAHDI, 2013).

The need to extract textual information encouraged the development of automated tools for reading documents, pattern recognition and knowledge extraction, the latter being the most important objective of text mining (PINTO et al., 2014). Text mining is similar to data mining, the main difference is in the organization of the structure, as the textual information is available in semi-structured or unstructured formats (KAUSHIK; NAITHANI, 2016).

This study has as scope the proposal of a method of information retrieval based on the use of computational tools that help in the visualization, organization and selection of relevant material for a scientific research in its initial phase. The research has as the theme the existing approaches to extracting data from scientific articles.

METHODOLOGY

A basis: *Web of Science* (WoS), was defined as the database to be used in this work, provided by Thomson Reuters is considered one of the most important bibliometric databases (YI et al., 2017). WoS also has resources that help filter searches, such as "Results Analysis" and "Citation Report", indispensable in the assertive search process.

In the initial process of the research, a technique was developed for the formation of the search string, a means of research in WoS, it must have well-defined boundaries to obtain a solid result. "Text mining" was defined as the main keyword, performing a search for the term in WoS, which resulted in a list of 3449 articles, where their metadata were exported.

The data obtained were entered into the VOSviewer® software, which enables the relationship between the records. A map was created based on the co-occurrence analysis of the authors' keywords, enabling the visualization of the links between them. A minimum occurrence of 30 times the keyword in each record was used as a criterion, in order to guarantee the relationship with the subject of the same. The result of the software can be seen in figure 1.

Based on the result exposed by the network analysis software, it was necessary to seek an understanding of the meaning of each item and the relationship between the items. After this process, the keywords were selected: "*machine learning*", "natural language processing", "information extraction" and "*knowlodge discovery*", confirming the strong relationship of the set with the main item "text mining". On the map, the colors serve to identify the various groups of keywords, and the place where each keyword appears indicates how close the items are to each other. The size of each circle indicates the number of times a keyword appeared in the set of articles

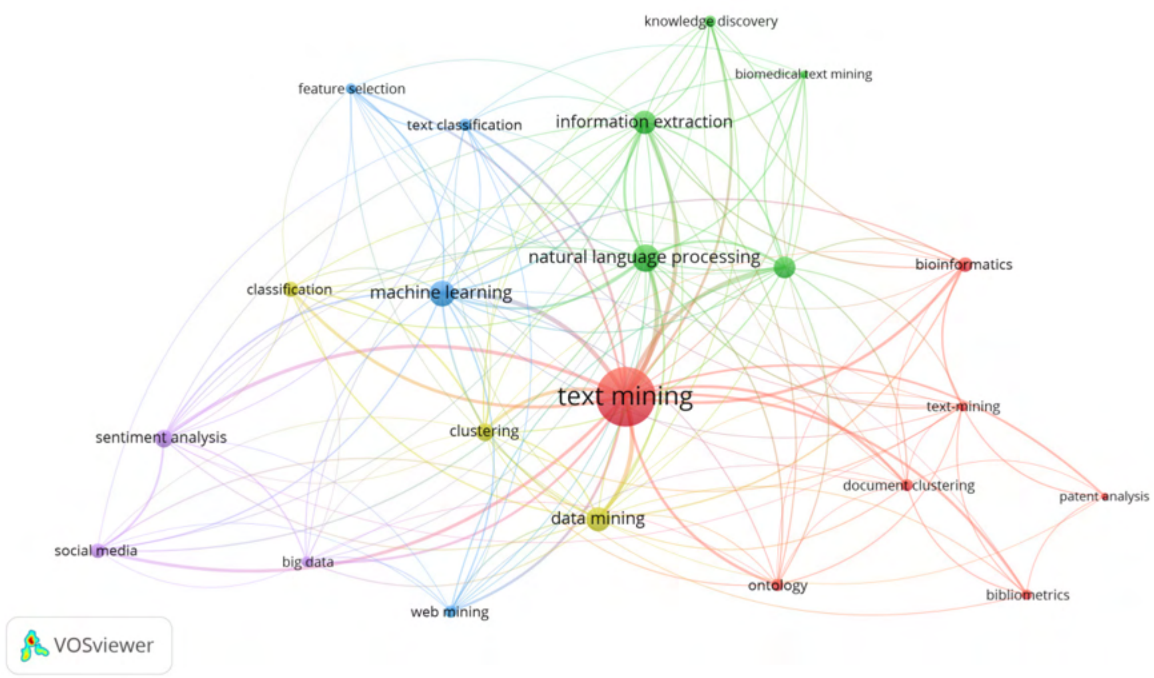


FIGURE 1 – VOSviewer® keyword co-occurrence map.

Source: Authors.

evaluated. The relevance of a keyword can be evaluated by the diameter of the circle that represents it.

After identifying the terms referring to the research topic, the search string was created, testing adherence to the researched topic and the range of results. Boolean operators were used to associate the keywords so that the search results must obtain the main term and at least one of the other defined terms, containing at least two of the five keywords. Still in the search string, a time delimitation of the last ten years was defined, thus obtaining the following string: "TS ("text mining" and ("machine learning" or "information extraction" or "natural language processing" " or "knowledge discovery")) and PY (2008-2017))".

Applying the string in the advanced search in the WoS base, a list with 776 records was obtained, distributed in several areas of knowledge. The "Results Analysis" feature of the database was then used to verify the distribution of the items by the categories in which they fit. The field "Web of Science Categories" presented a total of 63 categories, an analysis was carried out of these and of the areas of knowledge in which each one is found, in the end all categories little related to the research topic were excluded.

After applying the exclusion, the results were limited to the categories listed in Table 1, leaving a total of 247 articles in 104 journals, which were exported from WoS.

CATEGORY	QUANTITY
Computer Science, Information Systems	122
Computer Science, Artificial Intelligence	91
Information Science & Library Science	55
Computer Science, Theory & Methods	25
Computer Science, Software Engineering	25
Multidisciplinary Sciences	23
Computer Science, Interdisciplinary Applications	22
Operations Research & Management Science	16

Table 1 - Categories selected in the "Results Analysis" of the base: Web of Science.

Source: Authors.

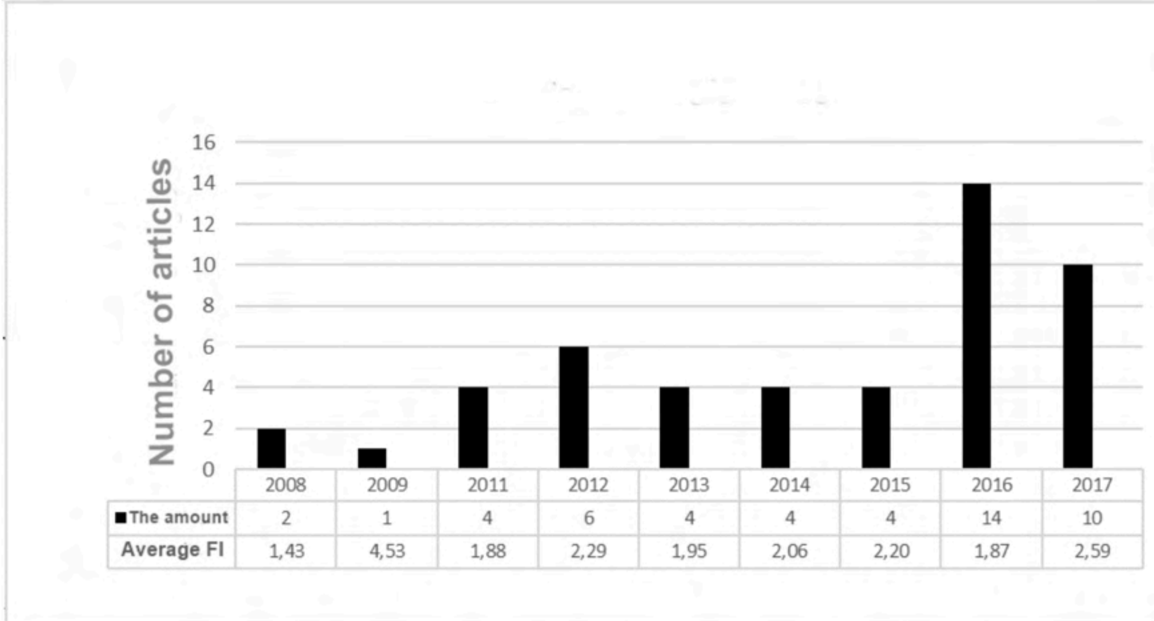
The results were used to feed an electronic spreadsheet and organized for analysis and detailing of the attributes of each record. The data provided by WoS has a total of 66 attributes, after filtering and treating the attributes, a spreadsheet was created with the following fields: Authors, Title, Magazine, Keywords, Keywords Plus, Abstract, Country, ISSN, Year of publication, DOI, impact factor and Qualis Capes.

The final step was the qualification of the results, where criteria were considered to define the importance of the article for the study, observing the research methods, the impact factor, the scope and availability of information. For this, it was necessary to read the title, abstract, keywords and in some cases introduction and conclusion, in addition to complementary research of information about the article and the journal.

RESULTS AND DISCUSSIONS

The search results allowed the identification of 49 articles, from 2008 to 2017, from 33 journals, with an average impact factor of 2.193.

Figure 2 shows the distribution of articles in the period (49% are recent articles and were published from 2016 onwards).



Average FI = average impact factor.

FIGURE 2: Annual distribution of articles selected from the database:Web of Science. Source: the authors.

Source: authors.

The articles were classified according to Qualis Capes in three evaluation areas: Engineering III, Computing and Multidisciplinary with, respectively, 36.7%, 61.2% and 36.7% of works published in the upper stratum (A1, A2 and B1), as shown in Table 2.

Assessment area	Qualis CAPES		
	A1	A2	B1
Engineerings III	8	7	3
Computation	18	7	10
Multi-subjects	3	10	5

TABLE 2 - Number of Articles in the Evaluation Areas and Qualis CAPES (Upper Stratum).

Source: The authors.

Appendix A displays the list of articles found, with author data, year, article theme based in the abstract, impact factor, Qualis Capes in the selected evaluation areas and the journal's ISSN. The compilation of this material made it possible to map the state of the art involving the mining of texts and related subjects.

TEXT MINING

The decision-making process involves steps that become extremely difficult when depending on the volume of textual information available on the Internet, making both the processing of results from a single search and the comparison of several searches impracticable (ABRAHAMS; BARKHI, 2013). To understand textual information in large quantities, it is necessary to use text mining, which comprises an area of study that deals with the construction of models and patterns of text resources, classification and sentiment analysis (PEROVSEK et al., 2016), (TROVATI et al., 2017).

Text mining is based on the use of data mining techniques, natural language processing, machine learning, information extraction and knowledge management (PEROVSEK et al., 2016). As the knowledge presented in numerous documents has an unstructured form (KAUSHIK; NAITHANI, 2016), it requires considerable efforts to determine which knowledge items are included, in this context an automatic recognition tool is essential (WANG et al., 2008).

Text mining can be described in the elementary form (KAUSHIK; NAITHANI, 2016), it works with information extraction and text classification techniques for the construction of machine learning using techniques such as: *Clustering* (JALIL *et al.*, 2016)(ATKINSON *et al.*, 2014) (CASAMAYOR; GODOY; CAMPO, 2012) (KAUSHIK; NAITHANI, 2016), heuristic methods (ABULAISH; ANWAR, 2012) (ATKINSON; FERREIRA; ARAVENA, 2009)(NOVACEK, VIT; BURNS, 2014), complex networks (ISAEVA; SUVOROVA; BAKHTIN, 2016), *Named Entity Recognition* (HASSANZADEH; KEYVANPOUR, 2013), *Singular Value Decomposition* (ABDUL-RAHMAN *et al.*, 2016) (ATKINSON *et al.*, 2014) and *Support Vector Machine* (WONG, 2012).

MACHINE LEARNING

The automatic evaluation of textual information can be defined as a classification approach, having location criteria in the text, which are mapped to perform a document score (MEHMOOD et al., 2017). This process defined as machine learning comprises several fields of study, such as logic, probability and statistics, combinatorial optimization and artificial intelligence (LEE et al., 2014).

Machine learning uses the approach of learning from data, it can be supervised, semi-supervised or unsupervised, which defines whether learning requires human support in training the technique (HASSANZADEH; KEYVANPOUR, 2013).

Machine learning techniques are used in conjunction with text mining in order to reveal concepts and connections, discover trends and associations between textual information (ROCHA, ROCIO; COBO, 2011) using high computational power. Analyzing machine learning algorithms, several solutions were developed, such as: *Support Vector Machine* (WONG, 2012) (TEICH *et al.*, 2016), *Bayesian Networks* (TROVATI *et al.*, 2017)(ATKINSON *et al.*, 2014), *Random Forest* (MEHMOOD *et al.*, 2017), *Nearest Neighbors methods* (GADRI; MOUSSAOUI, 2017)(ZHANG *et al.*, 2017), *Neural Network* (MORENO; REDONDO, 2016)(ISAEVA; SUVOROVA; BAKHTIN, 2016), *Fuzzy Set* (NOVACEK, VIT; BURNS, 2014) and *Ontology* (ISSERTIAL; TSUJI, 2015)(PROTAZIUK; LEWANDOWSKI; BEMBENIK, 2016)(CASAMAYOR; GODOY; CAMPO, 2012)(CONDE *et al.*, 2016).

These techniques have been applied to solve problems in several areas and this field continues to develop, according to (LEE et al., 2014) most machine learning studies approach the construction of new algorithms or applications in new areas, applying the knowledge in machine learning and classification techniques to extract information from textual data sets (ATKINSON et al., 2014).

NATURAL LANGUAGE PROCESSING

One of the techniques for automatic information extraction is Natural Language Processing (NLP), which performs a detailed analysis on unstructured textual information, often to satisfy specific information needs or a specific answer to a question (WANG et al., 2008).

For (ATKINSON; FERREIRA; ARAVENA, 2009) the application of NLP techniques is a key issue to extract relevant information from documents, considering this step as the predecessor for text mining based on complex language.

In text mining and NLP, the vector representation for a document, known as a frequency term, is common, having the frequencies of the terms contained in the document (KIM, 2016). This representation is essential in the use of document summarization techniques (YAO; WAN; XIAO, 2017).

Automated content-based text assignment, known as text categorization, is a type of supervised learning (ZHU; WONG, 2017) and has been applied in studies of language identification (TORNEY; YEARWOOD, 2012), information retrieval (WONG, 2012) opinion mining (STEINBERGER, 2012), and strongly related to Clustering techniques (RAFI et al., 2016)(MARX; DAGAN; SHAMIR, 2011)(WANG et al., 2008)(CASAMAYOR; GODOY; CAMPO, 2012)(TEICH et al., 2016)

INFORMATION EXTRACTION

The manual document classification process is an expensive process and may have inconsistencies, due to human interaction (MEHMOOD et al., 2017). The automatic document classification process is part of knowledge discovery, the main objective of text mining (TALIB et al., 2016). The justification for the classification is the reduction of information diversity, causing them to be grouped by similarity, avoiding information overload (ROCHA, ROCIO; COBO, 2011).

The most common techniques require that the text structures use the same vocabulary, so that it is possible to relate them, this way the words are treated as if they were independent (HUANG et al., 2012), making the meaning of the set not be understood. The extraction of the idea requires more complex representations of the content, allowing reasoning about the content (SAINT-DIZIER; MOENS, 2011) (MATTHIES, BENJAMIN; CONERS, 2017).

The problem of extracting information from a document is to find which terms best represent its scope (HADDOUD; ABDEDDAIM, 2015). How terms are selected significantly affects the accuracy of the extraction method.

Some complications are raised in extracting information, such as language ambiguity (TROVATI et al., 2017), (SONG; KIM; KIM, 2015)(LI; SUN; DATTA, 2013) and information redundancy (WANG et al., 2011)(GAMBHIR; GUPTA, 2016). Scientific writing is different from everyday writing, it employs structures and semantics designed to formulate and organize knowledge (LUO et al., 2016), such as making a hypothesis, analyzing data and drawing scientific conclusions (MOOHEBAT et al., 2014).

DISCOVERY OF KNOWLEDGE IN THE SCIENTIFIC AREA

In the scientific area, especially when it comes to bibliometrics, researchers have the need to find the maximum possible number of relevant publications in their research (ZHU; YAN; SONG, 2016). However, in academic database searches, the results present many articles irrelevant to the scope (MEHMOOD et al., 2017) and the classification of these is done manually to proceed with the search.

In the field of scientific research, there are already works with methodologies applied to the automatic extraction of information, such as: *Support Vector Machine*, *Hidden Markov Models* and *Conditional Random Fields* (HASSANZADEH; KEYVANPOUR, 2013), used in the identification of bibliometric attributes. The models have similarities in terms of identification, being necessary to label each attribute based on the recognition of textual patterns.

The recognition of technical terms are essential for understanding the techniques used in scientific research documents, (FAN; CHANG, 2008) uses the method: *Automatic Term Regognition* for the discovery of terminology in large textual volumes; (HADDOUD; ABDEDDAIM, 2015) developed a supervised learning system based on the association of the DPM-index and 18 other statistical characteristics for the extraction of key phrases in scientific documents; (JOORABCHI, ARASH; MAHDI, 2013) implements the classification of key phrases based on genetic algorithms without the need for manual feature selection, surpassing the classification by machine learning in general; (MOOHEBAT et al., 2014) proposes a new classification method focusing on the vocabulary of scientific articles and the differences between articles indexed in the IS (*Institute for Scientific Information*) and not indexed; (LEE et al., 2014) analyzes scientific publications through the application of the following methods:*Dirichlet-Multinomial Regression* and *Latent Dirichlet Allocation* to understand the relationship between the trend and the research capacity of universities; (RABIEI; HOSSEINI-MOTLAGH; HAERI, 2017) analyzes the interactions of researchers from environmental areas based on scientific publications in a database seeking to identify research priorities on this topic; (YI et al., 2017) uses the method: *Latent Dirichlet Allocation* to define the research scenario of a database detecting the changes in focus in a period of 25 years, tracing the subjects in evolutionary routes.

CONCLUSION

The methodologies used for scientific research have subjectivity and can be inefficient, as research, in most cases, does not use techniques or tools to support the researcher. A problem that is currently being studied, with a range of techniques, applications and methods for extracting and classifying data, working together to correlate the sentiment of textual information.

The proximity of the identification of relevance related to the search for scientific documents, brings significant contributions to researchers and their investigation strategies.

As a result of the study, it can be said that the method supported the process of research and analysis of articles, enabling greater visibility of items such as scope, authors, journal and the relationship between them.

The contributions of this work are consolidated both in the reduction of time to find relevant information in the midst of a large amount of documents, as in the ease of analyzing their content. It seeks to simplify the complex and manual processes in the retrieval of information in the scientific environment.

REFERENCES

ABDUL-RAHMAN, A.; ROE, G.; OLSEN, M.; GLADSTONE, C.; WHALING, R.; CRONK, N.; MORRISSEY, R.; CHEN, M. Constructive Visual Analytics for Text Similarity Detection. **Computer Graphics Forum**, [s. l.], v. 00, n. 00, p. 1–12, 2016.

ABRAHAMS, A. S.; BARKHI, R. Concept comparison engines : A new frontier of search. **Decision Support Systems**, [s. l.], v. 54, n. 2, p. 904–918, 2013. Disponível em: <http://dx.doi.org/10.1016/j.dss.2012.09.014>

ABULAISH, M.; ANWAR, T. A supervised learning approach for automatic keyphrase extraction. **International Journal of Innovative Computing, Information and Control**, [s. l.], v. 8, n. 11, p. 7579–7601, 2012.

ATKINSON, J.; FERREIRA, A.; ARAVENA, E. Discovering implicit intention-level knowledge from natural-language texts. **Knowledge-Based Systems**, [s. l.], v. 22, n. 7, p. 502–508, 2009. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2008.10.007>

ATKINSON, J.; GONZALEZ, A.; MUNOZ, M.; ASTUDILLO, H. Web metadata extraction and semantic indexing for learning objects extraction. **Applied Intelligence**, [s. l.], 2014.

CASAMAYOR, A.; GODOY, D.; CAMPO, M. Functional grouping of natural language requirements for assistance in architectural software design. **Knowledge-Based Systems**, [s. l.], v. 30, p. 78–86, 2012. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2011.12.009>

CONDE, A.; LARRANAGA, M.; ARRUARTE, A.; ELORRIAGA, J. A.; ROTH, D. LiTeWi : A Combined Term Extraction and Entity Linking Method for Eliciting Educational Ontologies From Textbooks. **JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY**, [s. l.], v. 67, n. 2, p. 380–399, 2016.

FAN, T.; CHANG, C. Exploring Evolutionary Technical Trends From Academic Research Papers. **Journal Of Information Science And Engineering**, [s. l.], p. 574–581, 2008.

GADRI, S.; MOUSSAOUI, A. Application of a new set of pseudo-distance in documents categorization. **Neural Network World**, [s. l.], p. 231–245, 2017.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques : a survey. **Artificial Intelligence Review**, [s. l.], 2016.

HADDOUD, M.; ABDEDDAIM, S. Accurate keyphrase extraction by discriminating overlapping phrases. **Journal of Information Science**, [s. l.], 2015.

HASSANZADEH, H.; KEYVANPOUR, M. A two-phase hybrid of semi-supervised and active learning approach for sequence labeling. **Intelligent Data Analysis**, [s. l.], v. 17, p. 251–270, 2013.

HUANG, L.; MILNE, D.; FRANK, E.; WITTEN, I. H. Learning a Concept-Based Document Similarity Measure. **Journal Of The American Society For Information Science And Technology**, [s. l.], 2012.

ISAEVA, E. V; SUVOROVA, V. A.; BAKHTIN, V. V. Supervized Machine Learning : Computer-Aided Development of a Specialized Dictionary. **Automatic Documentation And Mathematical Linguistics**, [s. l.], v. 50, n. 3, p. 104–111, 2016.

ISSERTIAL, L.; TSUJI, H. Information Extraction for Call for Paper. **International Journal Of Knowledge And Systems Science**, [s. l.], v. 6, n. December, 2015.

JALIL, A. M.; HAFIDI, I.; ALAMI, L.; KHOURIBGA, E. Comparative Study of Clustering Algorithms in Text Mining Context. **International Journal Of Interactive Multimedia And Artificial Intelligence**, [s. l.], v. 3, p. 42–45, 2016.

JOORABCHI, ARASH; MAHDI, A. E. Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. **Journal Of Information Science**, [s. l.], 2013.

KAUSHIK, A.; NAITHANI, S. A Comprehensive Study of Text Mining Approach. **Internacional Journal of Computer Science and Network Security**, [s. l.], v. 16, n. 2, p. 69–76, 2016.

KIM, M. Document Summarization via Convex-Concave Programming. **International Journal Of Fuzzy Logic And Intelligent Systems**, [s. l.], v. 16, n. 4, p. 293–298, 2016.

LEE, H.; KWAK, J.; SONG, M.; OUK, C. Coherence analysis of research and education using topic modeling. **Scientometrics**, [s. l.], 2014.

LI, C.; SUN, A.; DATTA, A. TSDW : Two-Stage Word Sense Disambiguation Using Wikipedia. **Journal Of The American Society For Information Science And Technology**, [s. l.], v. 64, n. 6, p. 1203–1223, 2013.

LUO, X.; XUAN, J.; LU, J. I. E.; ZHANG, G. Measuring the Semantic Uncertainty of News Events for Evolution Potencial Estimation. **Acm Transactions On Information Systems**, [s. l.], v. 34, n. 4, 2016.

MARX, Z.; DAGAN, I.; SHAMIR, E. Cross-partition clustering : revealing corresponding themes across related datasets. **Journal Of Experimental & Theoretical Artificial Intelligence**, [s. l.], v. 23, n. 2, p. 153–180, 2011.

MATTHIES, BENJAMIN; CONERS, A. Document Selection for Knowledge Discovery in Texts: Framework Development and Demonstration. **Journal Of Information & Knowledge Management**, [s. l.], v. 16, n. 4, p. 1–24, 2017.

MEHMOOD, A.; ON, B.; LEE, I.; CHOI, G. S. Prognosis Essay Scoring and Article Relevancy Using Multi-Text Features and Machine Learning. **Symmetry-Basel**, [s. l.], p. 1–16, 2017.

MOOHEBAT, M.; RAJ, R. G.; BINTI, S.; KAREEM, A. Identifying ISI-Indexed Articles by Their Lexical Usage : A Text Analysis Approach. **Journal Of The Association For Information Science And Technology**, [s. l.], 2014.

MORENO, A.; REDONDO, T. Text Analytics : the convergence of Big Data and Artificial Intelligence. **International Journal Of Interactive Multimedia And Artificial Intelligence**, [s. l.], p. 57–64, 2016.

NOVACEK, VIT; BURNS, G. A. P. C. SKIMMR : facilitating knowledge discovery in life sciences by machine-aided skim reading. **PeerJ**, [s. l.], p. 1–38, 2014.

PEROVSEK, M.; KRANJC, J.; ERJAVEC, T.; CESTNIK, B.; LAVRAC, N. TextFlows: A visual programming platform for text mining and natural language processing. **Science Of Computer Programming**, [s. l.], n. January, 2016.

PINTO, D.; GÓMEZ-ADORNO, H.; VILARIÑO, D.; SINGH, V. K. A graph-based multi-level linguistic representation for document understanding. **Pattern Recognition Letters**, [s. l.], v. 41, p. 93–102, 2014. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2013.12.004>

PROTAZIUK, G.; LEWANDOWSKI, J.; BEMBENIK, R. SAUText - a system for analysis of unstructured textual data. **Journal Of Intelligent Information Systems**, [s. l.], p. 369–389, 2016.

RABIEI, M.; HOSSEINI-MOTLAGH, S.-M.; HAERI, A. Using text mining techniques for identifying research gaps and priorities : a case study of the environmental science in Iran. **Scientometrics**, [s. l.], v. 110, n. 2, p. 815–842, 2017.

RAFI, M.; SHARIF, M. N.; ARSHAD, W.; RAFAY, H.; MOHSIN, S.; SHAIKH, M. S. Exploiting Document Level Semantics in Document Clustering. **International Journal Of Advanced Computer Science And Applications**, [s. l.], v. 7, n. 6, 2016.

ROCHA, ROCIO; COBO, A. Feature selection strategies for automated classification of digital media content. **Journal Of Information Science**, [s. l.], 2011.

SAINT-DIZIER, P.; MOENS, M. Knowledge and reasoning for question answering : Research perspectives. **Information Processing and Management**, [s. l.], v. 47, n. 6, p. 899–906, 2011. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2011.04.001>

SONG, M.; KIM, E. H.; KIM, H. J. Exploring author name disambiguation on PubMed-scale. **Journal of Informetrics**, [s. l.], v. 9, n. 4, p. 924–941, 2015. Disponível em: <http://dx.doi.org/10.1016/j.joi.2015.08.004>

STEINBERGER, R. A survey of methods to ease the development of highly multilingual text mining applications. **Language Resources and Evaluation**, [s. l.], p. 155–176, 2012.

TALIB, R.; HANIF, M. K.; AYESHA, S.; FATIMA, F. Text Mining : Techniques , Applications and Issues. **International Journal Of Advanced Computer Science And Applications**, [s. l.], v. 7, n. 11, p. 414–418, 2016.

TEICH, E.; DEGAETANO-ORTLIEB, S.; FANKHAUSER, P.; KERMES, H.; LAPSHINOVA-KOLTUNSKI, E. The Linguistic Construal of Disciplinarity : A Data-Mining Approach Using Register Features. **Journal Of The Association For Information Science And Technology**, [s. l.], v. 67, n. 7, p. 1668–1678, 2016.

TORNEY, R.; YEARWOOD, J. Using Psycholinguistic Features for Profiling first language of authors. **Journal Of The American Society For Information Science And Technology**, [s. l.], v. 63, n. 6, p. 1256–1269, 2012.

TROVATI, M.; HAYES, J.; PALMIERI, F.; BESSIS, N. Automated Extraction of Fragments of Bayesian Networks from Textual Sources. **Applied Soft Computing**, [s. l.], 2017. Disponível em: <http://dx.doi.org/10.1016/j.asoc.2017.07.009>

WANG, L.; FUKETA, M.; MORITA, K.; AOE, J. Context constraint disambiguation of word semantics by field association schemes. **Information Processing and Management**, [s. l.], v. 47, n. 4, p. 560–574, 2011. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2011.01.001>

WANG, W. M.; CHEUNG, C. F.; LEE, W. B.; KWOK, S. K. Mining knowledge from natural language texts using fuzzy associated concept mapping. **Information Processing & Management**, [s. l.], v. 44, p. 1707–1719, 2008.

WONG, T. Learning to adapt cross language information extraction wrapper. **Applied Intelligence**, [s. l.], n. June 2011, p. 918–931, 2012.

YAO, J.; WAN, X.; XIAO, J. Recent advances in document summarization. **Knowledge and Information Systems**, [s. l.], 2017.

YI, Z.; CHEN, H.; JIE, L.; GUANGQUAN, Z. Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. **Knowledge-Based Systems**, [s. l.], 2017. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2017.07.011>

ZHANG, X.; ZHANG, X.; LIU, H.; LIU, X. Multi-task clustering through instances transfer. **Neurocomputing**, [s. l.], v. 251, p. 145–155, 2017.

ZHU, D.; WONG, K. W. An evaluation study on text categorization using automatically generated labeled dataset. **Neurocomputing**, [s. l.], 2017. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2016.04.072>

ZHU, Y.; YAN, E.; SONG, M. Understanding the evolving academic landscape of library and information science through faculty hiring data. **Scientometrics**, [s. l.], 2016.

APPENDIX A - LIST OF SELECTED ARTICLES – BASE DA *WEB OF SCIENCE*

Authors	Year	Themes	FI	QUALIS CAPES E.III	COMP	MULT	ISSN of periodical
Fan & Chang	2008	Trends in evolutionary techniques	0,468	--------	--	---------	1016 2364
Wangt et al.	2008	knowledge mining	2,391	----	A1	----------	03
Atkson et al.	2009	Automatic discovery of implicit rhetorical information	4,529	A1	A1	A1	
Marx et al.	2011	Cross partition clustering	1,372				
Rocha & Cobo	2011	Strategies for automated sorting					
Saint -Dizier & Moens	2011	Research perspectives on knowledge and reasoning					
Wang et al	2011	Formal Disambiguation					
Abulaish &| Anwar	2012	Identification of key phrases					
Casamayor et al	2012	Mining and clustering functionality					
Huang et al	2012	Concept-based document simularity measure					
steinberger	2012	A survey of highly multilingual methods					
Torney et al	2012	Profiling the authors' first language					
Wong	2012	Information Extraction					
Abrahams & Barkhi	2013	Comparison engine as a decision support tool					
Hassandeh & keyvannpour	2013	Sequential labeling					
Joorabchi & Mahdi	2013	Machine learning-based key phrase annotation					
Li et al	2013	Disambiguation of the meaning of the word					
Atkinson et all	2014	Classification accuracy and metadata quality					
Leet et al	2014	automatic text parsing					
Novacek & Burns	2014	Discovery of knowledge in life sciences					
Pinto et all	2014	A graphic-based representation of textual documents					

Haddoud & Abdeddaim	2015	supervised learning system					
Issertial & Tsuji	2015	Extraction of information for publications					

AUTORES	ANO	TEMAS	FI	QUALIS CAPES			ISSN do Periódico
				E.III	COMP	MULT	
Fan & Chang	2008	Tendências das técnicas evolutivas	0,468	-	-	-	1016-2364
Wang *et al.*	2008	Mineração de conhecimento	2,391	-	A1	-	0306-4573
Atkinson *et al.*	2009	Descoberta automática de informação retórica implícita	4,529	A1	A1	A1	0950-7051
Marx *et al.*	2011	*Clustering* de partição cruzada	1,384	A2	-	-	0952-813X
Rocha & Cobo	2011	Estratégias para classificação automatizada	1,372	-	A2	B1	0165-5515
Saint-Dizier & Moens	2011	Perspectivas da pesquisa sobre conhecimento e raciocínio	2,391	-	A1	-	0306-4573
Wang *et al.*	2011	Desambiguação formal	2,391	-	A1	-	0306-4573
Abulaish & Anwar	2012	Identificação de frases-chave	1,667	B3	C	-	1349-4198
Casamayor *et al.*	2012	Funcionalidade de mineração e agrupamento	4,529	A1	A1	A1	0950-7051
Huang *et al.*	2012	Medida de similaridade de documentos baseada em conceitos	2,452	A1	A1	-	1532-2882
Steinberger	2012	Uma pesquisa de métodos altamente multilíngues	0,738	-	B1	-	1574-020X
Torney *et al.*	2012	Traçando o perfil do primeiro idioma dos autores	2,452	A1	A1	-	1532-2882
Wong	2012	Extração de informações	1,904	B1	B1	B1	0924-669X
Abrahams & Barkhi	2013	Mecanismo de comparação como uma ferramenta de suporte à decisão	3,222	A1	A1	-	0167-9236
Hassanzadeh & Keyvanpour	2013	Rotulagem sequencial	0,772	-	B1	-	1088-467X
Joorabchi & Mahdi	2013	Anotação de frases-chave baseada em aprendizado de máquina	1,372	-	A2	B1	0165-5515
Li *et al.*	2013	Desambiguação do sentido da palavra	2,452	A1	A1	-	1532-2882
Atkinson *et al.*	2014	Precisão de classificação e qualidade de metadados	1,904	B1	B1	B1	0924-669X
Lee *et al.*	2014	Análise automática de texto	2,147	A2	A1	A2	0138-9130
Novacek & Burns	2014	Descoberta de conhecimento em ciências da vida	2,177	A1	-	A2	2167-8359
Pinto *et al.*	2014	Uma representação baseada em gráficos de documentos textuais	1,995	B1	A1	A2	0167-8655
Haddoud & Abdeddaim	2015	Sistema de aprendizagem supervisionada	1,372	-	A2	B1	0165-5515
Issertial & Tsuji	2015	Extração de informação para publicações	-	-	-	-	1947-8208
Moohebat *et al.*	2015	Análise de texto para identificação de artigos indexados por isi	2,322	-	-	-	2330-1635
Song *et al.*	2015	Desambiguação do nome do autor	2,92	-	A2	A2	1751-1577
Abdul-Rahman *et al.*	2016	Detecção de similaridade de texto	1,611	-	A1	-	0167-7055
Conde *et al.*	2016	Extração de termo e vinculação de entidade	2,322	-	-	-	2330-1635
Isaeva *et al.*	2016	Sistema de aprendizagem supervisionada	-	-	-	-	0005-1055
Jalil *et al.*	2016	Bancos de dados de descoberta de conhecimento	-	-	C	B2	1989-1660
Kaushik & Naithani	2016	Revisão de técnicas de mineração de texto	-	-	C	-	1738-7906
Kim	2016	Côncava-convexa para sumarização de documentos	-	-	-	-	1598-2645
Luo *et al.*	2016	Medindo a incerteza semântica	2,312	-	-	-	1046-8188
Moreno & Redondo	2016	A convergência do big data e inteligência artificial	-	-	C	B2	1989-1660
Perovsek *et al.*	2016	Plataforma de programação visual para mineração de texto	1,064	-	A2	-	0167-6423
Protaziuk *et al.*	2016	Um sistema para análise de dados textuais não estruturados	1,294	-	B1	-	0925-9902
Rafi *et al.*	2016	Agrupamento de documentos em semântica de nível de documento	-	-	-	-	2158-107X
Talib *et al.*	2016	Técnicas, aplicativos e questões sobre mineração de texto	-	-	-	-	2158-107X
Teich *et al.*	2016	A interpretação linguística da disciplinaridade	2,322	-	-	-	2330-1635
Zhu *et al.*	2016	Ciência da informação por meio da faculdade de contratar dados	2,147	A2	A1	A2	0138-9130
Gadri & Moussaoui	2017	Categorização automática de texto com k-nn	0,394	-	-	-	1210-0552
Gambhir & Gupta	2017	Técnicas de Sumarização de Texto	2,627	-	A2	-	0269-2821
Matthies & Coners	2017	Seleção de documentos para descoberta de conhecimento	-	B3	-	-	0219-6492
Mehmood *et al.*	2017	Recursos multi-texto e aprendizado de máquina	1,457	-	-	-	2073-8994
Rabiei *et al.*	2017	Técnicas de mineração de texto para identificar lacunas de pesquisa	2,147	A2	A1	A2	0138-9130
Trovati *et al.*	2017	Extrair e construir fragmentos de redes bayesianas	3,541	A2	A1	A2	1568-4946
Yao *et al.*	2017	Sumarização de documentos	2,004	-	A2	A2	0219-1377
Zhang *et al.*	2017	Sistemas baseados em conhecimento de análise bibliométrica	3,317	A2	A1	A2	0925-2312
Zhang *et al.*	2017	*Clustering* multitarefa	4,529	A1	A1	A1	0950-7051
Zhu & Wong	2017	Estudo sobre categorização de textos	3,317	A2	A1	A2	0925-2312

FI= IMPACT FACTOR, E.III – ENGINEERING III, COMP = COMPUTING, MULT=MULTIDISCIPLINARY.