ICKM suceg 2019
KNOWLEDGE AND LEARNING GOVERNANCE FOR INNOVATION

IDENTIFICATION OF KEYWORDS BASED ON ARTIFICIAL INTELLIGENCE FOR KNOWLEDGE MANAGEMENT

Dildre Georgiana Vasques
Doctoral candidate, ICMC-USP, orcid.org/0000-0002-1339-4518, dildre.vasques@usp.br

Paulo Sérgio Martins Pedro
PhD, FT-UNICAMP, orcid.org/0000-0002-6303-5564, paulo@ft.unicamp.br

Solange Oliveira Rezende
Dr, ICMC-USP, orcid.org/0000-0002-5233-7639, solange@icmc.usp.br

SUMMARY

Goal: *In this work, from a set of textual scientific documents from public databases, we carried out an automatic analysis to detect the dominant terms in the field of Knowledge Governance.*

Design/Methodology/Approach: *We applied Text Mining techniques combined with Complex Networks metrics to identify the keywords and their relationships.*

Results: *As a result, we have shown that information and knowledge may be quickly and automatically extracted from scientific documents. This process may reduce costs (due to the automation) and allow the sharing and dissemination of information which would otherwise be difficult to accomplish manually. One can appreciate the value of such analysis by considering that it would take hours of intensive (painstaking and error-prone) human labor to achieve similar results.*

ORGANIZED BY
SPONSORS

Limitations of the research: *The initial search for the document set is still manual, and thus may take time. However, ongoing work is already addressing this issue.*

Originality/value: *We have found the terms that are central to the context of Knowledge Governance. The domain analysts can then use the context of application of these terms and combine them with their own knowledge to facilitate decision making and selection of future strategies. The analysis is superior to a simple counting of terms or to an extraction approach of relationships based on the co-occurrence of terms since it is implemented and supported by relatively sophisticated Artificial Intelligence techniques, e.g. Text Mining and specifically Association Rules. These techniques were combined with Complex Networks to produce the desired outcome.*

Keywords: *association rules, complex networks, decision making, knowledge governance, text mining.*

IDENTIFICAÇÃO DE PALAVRAS-CHAVE BASEADAS EM INTELIGÊNCIA ARTIFICIAL PARA GESTÃO DO CONHECIMENTO

RESUMO

Objetivo: *Neste trabalho, a partir de um conjunto de documentos científicos textuais de um banco de dados público, realizamos uma análise automática para detectar os termos dominantes no campo da Governança do Conhecimento.*

Design/Metodologia/Abordagem: *Aplicamos técnicas de Mineração de Texto combinadas com métricas de Redes Complexas para identificar as palavras-chave e suas relações.*

Resultados: *Como resultado, mostramos que a informação e o conhecimento podem ser obtidos rápida e automaticamente a partir de documentos científicos. Esse processo pode reduzir custos (devido à automação) e permitir a disseminação de informações que, de outra forma, seriam difíceis de realizar manualmente. Pode-se apreciar o valor de tal análise considerando que seriam necessárias horas de trabalho humano intensivo (penoso e propenso a erros) para alcançar resultados semelhantes.*

Limitações da pesquisa: *A busca inicial pela coleção de documentos ainda é um processo manual, e portanto pode requerer tempo. No entanto, já existe trabalho em andamento contemplando esta limitação.*

Originalidade/valor: *Encontramos os termos que são centrais para o contexto da Governança do Conhecimento. Os analistas de domínio podem então usar o contexto de aplicação desses termos e combiná-los com seus próprios conhecimentos para facilitar a tomada de decisões e seleção de estratégias futuras. Esta análise é superior a uma contagem simples de termos ou a uma abordagem de extração de relações baseadas em coocorrência destes termos, uma vez que é implementada e suportada por técnicas de Inteligência Artificial relativamente sofisticadas, ou seja, Mineração de Texto e especificamente Regras de Associação. Estas técnicas foram combinadas com Redes Complexas para produzir o resultado desejado.*

Palavras-chave: *regras de associação, redes complexas, tomada de decisão, governança do conhecimento, mineração de texto.*

1 INTRODUCTION

The access and integration of expertise knowledge represent one of the key capabilities that may lead organizations to profitability and competitive advantage (Grant, 1996). The development of these capabilities exclusively within a company is no longer enough to cope with technological evolution. Thus, firms can stimulate internal creativity and innovation by tapping external sources of knowledge (Ferraris, Santoro, & Dezi, 2017).

However, this type of knowledge is generally linked to intellectual property rights (patent). One way out of this problem lies in the public offering of technological knowledge generated by academic research (Antonelli, 2016). It is widely known, especially within the Knowledge Management (KM) community, that there is a large amount of information available in the form of public documents (e.g. in the Internet, academic papers, digital libraries and other repositories) that may be explored (Cleveland, 1998; Marwick, 2001; Cody et al., 2002; Stenmark, 2002; Mao, et al., 2016; Santoro et al., 2018). There is vital knowledge that can be acquired and assimilated from these sources of information.

Nevertheless, resources are usually lacking to fully and efficiently explore this wealth of information. The processing of the large collections of documents may be deemed cumbersome, time-consuming and a costly endeavor if they are to be manually executed (Beliga, Meštrović, & Martinčić-Ipšić, 2015). Thus, for companies to create and capture value from external sources (Hahsler & Karpienko, 2017), technological innovations must also be integrated with their own internal KM practices and process (Fong & Chen, 2011; Ashok, Narula, & Martinez-Noya, 2016).

The technological approaches that focus on the development and application of information and communication technology (ICT tools) tools in companies, along with people-centered approaches, make up two important categories of governance mechanisms that facilitate knowledge processes (Fong & Chen, 2011). It is known that the influence of organizational mechanisms on knowledge acquisition is mediated by the technological mechanisms (Fong & Chen, 2011), thus suggesting their essential role in KM in the internet age.

Within this context, the exploration of knowledge in public academic research documents mediated by technological mechanisms can be the basis for discoveries at a reduced cost, thus supporting KM and Knowledge Governance (KG) systems in organizations. In this sense, organizations may rely on the support of Artificial Intelligence (AI) techniques. Through this technology, the explicit knowledge available can be contextualized to produce new knowledge that

favors competitive market advantages (Chung, 2014; Bouakkaz et al., 2017; Liébana-Cabanillas, Marinković, & Kalinić, 2017; Rekik et al., 2018; Duan, Edwards, & Dwivedi, 2019). Work can be found in the literature that focus on the analysis of how humans and new AI can be complementary in organizational decision-making (Jarrahi, 2018; Daugherty & Wilson, 2018). AI is believed to help employees in the organization make better decisions by increasing their analytical skills as well as their creativity. According to Jarrahi (2018), one way to materialize the synergistic relationship between AI and humans is to combine the speed of AI in gathering and analyzing information with the superior intuitive insight and perception of humans.

In particular, KG is one of the fields that is evolving. The KG construct is a complex system that encompasses the global and public application levels. It spans the corporate and individual levels and focuses on the understanding and coordination of the knowledge processes (Pemsel & Muller, 2012). It has a large number of unexplored documents and it may benefit from the application of AI techniques while exploring these documents in the search for information and knowledge.

One way of exploring the knowledge contained in KG documents is the search and extraction for keywords (Beliga, Meštrović, & Martinčić-Ipšić, 2015). The keywords are the relevant terms that appear in a set of documents. Note that these are not necessarily the most frequent terms, but the ones that are most semantically connected with other terms. They form a network of ideas that compose the subject. Keyword Extraction (KE) is an automatic method used to identify a set of terms that synthesize the subject of either a single (Rose at al., 2010) or a whole collection of documents (Wu & Agogino, 2004; Berry & Kogan, 2010).

This task may be implemented with the support of Text Mining (TM, a sub-area of AI) techniques (Jones, 1999; Sebastiani, 2002; Berry & Kogan, 2010). The TM process allows the automatic analysis of a large set of texts with little manual intervention (Chen, 2001; Rezende et al., 2003), in order to find rules and patterns that result in useful knowledge (Cohen et al., 2010). In turn, the rules extracted by AI techniques can be represented and analyzed using the graph-based text representation approach. In this approach, the relationships between terms can be explored through different metrics extracted from graphs (Sonawane & Kulkarni, 2014).

Therefore, in this work, we propose the identification of keywords across a set of documents in the KG domain assisted by an AI technique, specifically Association Rules (a TM task), and Graph-based Text Representation, specifically Complex Networks metrics. The keywords may be

regarded as extracted information that can assist in the general understanding of the KG construct itself to increase the technical and innovative capacity of the organization. The keywords and their relationships may then be scrutinized by the domain analyst, who may identify the context where they appear and analyze their relevance for a scenario of decision-making and knowledge management.

The remainder of this article is organized as follows: Section 2 defines the theoretical context of our research. Section 3 outlines the methodology used by explaining how we perform the procedures for data analysis. In Section 4, we explore the findings and practical implications of our methodology. Finally, in Section 5, we discuss the contributions obtained, clarify some limitations of the approach and propose future research.

2 BACKGROUND AND LITERATURE REVIEW

In this section, we review the basic areas involved in this work, namely Knowledge Governance and Artificial Intelligence, Text Mining and Complex Networks.

2.1 RELATIONSHIP BETWEEN KNOWLEDGE GOVERNANCE AND ARTIFICIAL INTELLIGENCE

Within the realm of organizations, KG is a concept that focuses on understanding and coordinating the processes of use, sharing, integration and knowledge creation (Foss, Husted, & Michailova, 2010; Pemsel & Muller, 2012; de Sá Freire et al., 2017). However, this process is not trivial, as this approach has an interdisciplinary nature that spans the fields of KM, organizational learning, organization studies, strategy and human resource management (Foss, 2007; Michailova & Foss, 2009).

There is a network of organizational relationships that must be managed to strike a balance between dependency and power. This network of relationships includes interactions between macro (organizational) and micro (individual) levels, as well as inter and intra organizational relationships (people, processes, technologies) (Pemsel et al., 2014). In order to understand these relationships and the distribution of rights and responsibilities between the parties involved, KG has developed a set of principles and mechanisms (Monks & Minow, 1995; Grandori, 2001; Foss et al. 2010; Gooderham, Minbaeva, & Pedersen, 2011).

Governance mechanisms must be used as a strategic resource if organizations are to achieve successful innovation and business development (Spender, 1996; Pietersen, 2010; Fong & Chen, 2011). The sum of actual and potential resources provided by personal and organizational networks (processes and technologies) can be defined as social capital (Nahapiet & Ghoshal, 1998; Tsang, 2000). For an organization's social capital to be used for innovation, it must first be transformed into strategic learning. Therefore, KG mechanisms must be translated into business activities and processes (Fong & Chen, 2011). According to Antonelli (2016), in view of changing market conditions, organizations that value high-quality governance mechanisms tend to manifest creative behavior, increasing their technological and innovative levels.

Currently, KG's nature is undergoing a change, or rather an evolution, due to the technological advancement and the emergence of Web 2.0. Technologies that support KM are faced with the need to include new AI technologies used to assist knowledge acquisition and support decision making (Becerra-Fernandez & Sabherwal, 2014). Thus, it is necessary to recognize the interrelationships in the governance mechanisms management and the new AI technologies in KM, thus seeking a proper balance between the use of these technologies and the social mechanisms (Fong & Chen, 2011).

Knowledge is recognized as a core competency, a primary source of competitive advantage and value creation for organizations worldwide (Liu et al., 2018). Like KM and human intelligence, AI is also associated with knowledge (Becerra-Fernandez & Sabherwal, 2014). Through algorithmic models, AI techniques enable computers to perform tasks that resemble the human thinking ability (Becerra-Fernandez & Sabherwal, 2014; Rekik et al., 2018). The terms associated with technological development related to the application of AI in business and management have changed over time. Early systems were called expert systems, replaced in the early 21st century by the term knowledge-based systems. Currently, the most common terms are machine learning and data mining (Becerra-Fernandez & Sabherwal, 2014).

Several papers report that AI is transforming business with the support of data manipulation and at a reduced cost (Duan et al., 2019; Daugherty & Wilson, 2018). Data mining technologies can be used as strategies to activate KM systems in organizations because they are capable of processing large amounts of information beyond human capacity. For Jarrahi (2018), AI and other intelligent technologies can help human decision makers by providing different types of analysis

and by identifying relationships between many factors. Thus, these technologies allow human decision makers to act on new sets of information, thus enabling the generation of new ideas.

2.2 TEXT MINING

Knowledge Discovery in Databases (KDD) is the process of finding and interpreting patterns from data (Fayyad et al., 1996; Becerra-Fernandez & Sabherwal, 2014). Within KDD we have: 1) Data Mining (DM), which extracts information from relational (structured) databases, and 2) Text Mining (TM), which tackles mining written texts in natural language, thus dealing with unstructured databases. The analysis of textual data requires context analysis (Bhardwaj & Khosla, 2017). Learning how to mine texts can lead to a significant amount of new knowledge. The estimate is that 80 % of the world's online content is available in textual format such as HTML documents and files, emails, blogs, newspaper articles, etc. (Chen, 2001).

Academic, public domain technological knowledge can also be found on the Web. Therefore, business organizations may profit from the mining of this type of document. For the texts to be processed, they must first go through the structuring (pre-processing) phase, using linguistic techniques combined with statistical ones. After this initial phase, TM technology enables the discovery of patterns for the performance of quantitative or qualitative analysis in textual documents (Aranha & Passos, 2006; Hashimi, Hafez, & Mathkour, 2015).

The analysis based on the extraction of Association Rules is one of the TM tasks. It is used to find patterns that describe the highly associated characteristics within the data (Tan, Steinbach, & Kumar, 2019). The mining of Association Rules is a DM-originated task that emerged in the context of retail applications. It aims at the analysis of the so-called "shopping baskets" of customers (Agrawal, Imielinski, & Swami, 1993; Tan, Steinbach, & Kumar, 2019).

When this task is applied to texts, the Association Rules explicit the words that are most often related by extracting relations of type A→B, meaning that when a term A occurs, a term B also tends to occur. To generate the Association Rules from the texts, the documents are first mapped onto transactions (set of words that appear together in a certain textual space). Mapping can be based on sentences, paragraphs, or a sliding window. The sentence mapping considers the words that appear between the punctuation marks that indicate phrase term. The paragraph mapping considers the set of all words that are contained between the punctuation marks that indicate phrase ending, followed by a line break. Sliding window mapping does not consider punctuation, as it is

based on a user-stipulated maximum distance limit, where words (items) should appear together in the text (Rossi & Rezende, 2011).

After performing the document-mapping step, we can extract the Association Rules. The most common measures used in extracting association rules are support and confidence. Support determines how often a rule (A→B) occurs in a specific data set, whereas confidence determines how often A→B occurs together in relation to the total number of transactions where A occurs (Hahsler & Karpienko, 2017; Bouakkaz et al., 2018).

Some tools that extract Association Rules provide the option of selecting the automatic calculation of the minimum support, taking into account the average frequency of the words in the transactions. The automatic calculation exempts the user from knowing the characteristics of the document or the collection of documents (Rossi & Rezende, 2011). The assignment of a minimum value to support is a complex task, because if this value is too high, the rules that include low frequency items are not found and, on the other hand, if this value is too low, a large number of rules is generated (simple co-occurrence).

2.3 COMPLEX NETWORK

Network construction techniques (graphs) make it possible to extract information from structured, semi-structured or even unstructured data sources (Jiang et al., 2010). When applied to texts, this model is able to represent relationships between documents or between terms that make up the documents themselves.

A network can be described as a graph and therefore inherits its conceptual properties (Beliga & Martinčić-Ipšić, 2015). A graph is mathematically defined as a structure composed of two sets: V (vertices) and E (edges), resulting in the pair G = (V; E). When a link joins two vertices (nodes) it means that there is an adjacency between them. There are two types of graphs: directed and non-directed. In a directed graph, also called a digraph, the edges have orientations, thus representing a non-symmetric relationship between the two objects. In a non-directed graph, the relation between the vertices are symmetric. Directed edges are used to indicate the order in which nodes (words) appear in the text. If there is no interest from the researcher to maintain order in the mutual interactions between nodes, the undirected edges may be used to connect the related nodes (Chang & Kim, 2013).

A network representation can increase the number of links that make up the intermediate path between different words, thus allowing the extraction of all the words involved in a relationship. This approach also provides a number of metrics to rank the terms according to different criteria, e.g. the degree of a node, clustering coefficient and betweenness centrality.

2.4 RELATED WORK

According to Zanin et al. (2016), the scientific community has been using DM techniques (including TM) in conjunction with Complex Network Theory to extract information from large datasets. This is because both share the same overall goal of extraction of information from complex systems. We searched for literature that involved the joint application of the TM approach based on Association Rules and the representation of graph texts (or networks) for the extraction of information and/or knowledge.

Hahsler and Karpienko (2017) argue that graph-based techniques make it possible to view Association Rules using vertices and edges, where vertices typically represent items (words) or sets of items and edges indicate relationships in terms of rules. The authors themselves have presented an interactive visualization method that allows the exploration and interpretation of highly complex scenarios. According to the authors, the method can be used to analyze large sets of Association Rules using the R software for statistical computing.

Some authors have proposed the extraction of Association Rules from the graph itself, such as Fan et al. (2015), who extracted the rules from the graph standard (GPARs) for social media marketing. The authors extended the Association Rules to sets of items to uncover regularities between entities on social graphs and identify potential customers by exploiting social influence. Berlingerio et al. (2009) studied time-dependent Association Rules through graph patterns, that is, graph evolution. For this, they adopted different semantics for the support and confidence values found in the graph itself.

Netzer et al. (2012) applied TM techniques combined with an automatic conditional random field approach (CRF) along with network analysis techniques to convert data extracted from web published texts into a semantic network. The goal of this application was to inform companies about the structure of the market and its relationships. However, the authors did not use Association Rules as the basis for network formation, but co-occurrence relations of terms to form semantic networks of the market structure. Similarly, Sonawane and Kulkarni (2014) also constructed a

graph based on the cooccurrence of words in the document to extract the relationships between terms. For this, the authors explored different metrics extracted from the graphs.

Regarding the specific search for keywords or keyphrases, we identified work such as Boudin (2013), who presented and compared several centrality measures for the extraction of graph-based keyphrases. Through experiments performed on three standard data sets from different languages and domains, the author shows that simple degree centrality achieves results comparable to the widely used TextRank algorithm in the area, and that proximity centrality achieves the best results in short documents.

Krapivin et al. (2010) used Natural Language Processing techniques to improve different machine learning approaches to the freely available keyphrase extraction problem of scientific articles. Lahiri, Choudhury, and Caragea (2014) experimented with a series of measures of centrality in word and noun networks and analyzed their performance in four reference data sets. They found that centrality measures perform as well as or better than PageRank (the algorithm used by Google's search engine to position websites) and are simpler (e.g. degree, strength and size of the neighborhood) to use for keyword and keyphrase extraction.

3 METHODOLOGY

Most academic publications are available in digital format. To assist in the process of extracting information (keywords and knowledge) from these texts, we may rely on TM techniques. In this sense, the present work proposes a process to extract simple and compound terms present in a textual corpus using Association Rules and Complex Network metrics. In addition to that, our work is also able to identify the relationships between these keywords, which also reveals knowledge.

The goal is to automatically select the most representative terms of a domain (keywords), thus synthesizing and identifying the concepts that support the arguments of that domain. The following steps aim to filter out and interconnect relevant content from the text set to support the knowledge discovery process:

Step 1 - Document Selection. The scope of this step is the selection of a set of documents to be used by the mining application. After this selection, the user must determine the parts of the text to be used. This decision must be made based on the focus of the problem to be addressed.

Depending on the purpose of mining, whole texts may be processed. In other cases, only a few sections are sufficient.

Step 2 - Association Rules Extraction. The Association Rules applied to texts are used to identify words that stand out by their frequency of occurrence, as well as (and mainly) more frequent relationships between two words. These relationships come in the form of the rule A→B, meaning that when term A occurs, term B also tends to occur. It is possible to measure the strength of an association rule in terms of support (how often a rule occurs in the transactions) and confidence (how often a rule occurs in relation to the total number of transactions where A occurs). The tasks required to extract the association rules are described below. In order to perform them, we use the open-access tool Features gEnerator (FEATuRE) (Rossi & Rezende, 2011):

a) **Preprocessing:** In this sub-step, the documents written in natural language should be passed to a format suitable for the mining process. Some data cleansing techniques are applied to improve processing time and the quality of extracted patterns. Data cleaning consists of the task of removing some special characters (digits, accents, and line breaks), in addition to stop words, i.e. words considered irrelevant (articles, prepositions, conjunctions).

b) **Mapping of textual documents in transactions:** After the preprocessing step, the text set is mapped onto transactions (different sets of items, i.e. words). In association analysis, a collection of zero or more items is called an item set. To obtain transactions, one can choose to map sentences, paragraphs, or sliding windows of different dimensions.

c) **Extraction of Association Rules:** For the extraction of Association Rules from the transactions defined in the previous sub-step, we can use the automatic minimum support, which considers the average frequency of the words in the transactions. Automatic support exempts the user from knowing the document's characteristics and avoids manually setting of a low minimum support value which could in turn generate too many trivial rules (i.e. simple cooccurrence). With the rules obtained, we may obtain a list based on how often simple terms and compound terms appear in the dataset. This list may serve as an indication of the terms that stand out in the domain and underpin the domain being investigated.

d) **Extracting frequent itemsets:** With the Association Rules obtained, we may use frequent itemsets to represent the characteristics of a document. When frequent itemsets are extracted from parts of a document rather than document collections, terms that are more meaningful will be obtained as they are related to specific document contexts, i.e. they are not simply a set of words scattered across documents in a collection. Frequent itemsets are sets of items that appear together in at least *n* % of the transactions, where *n* is the minimum support value. Consider the following transactions: {A B C}, {A C}, {C D}, {A B}, {B D}, {D}. If, for example, the support value is 1/3, it means that the itemsets must appear in at least two of the six transactions. Thus, {A B} is an example of frequent itemsets.

Step 3 - Create a Complex Network. The complex network approach can provide metrics that help us identify the most important network terms (i.e. keywords) and the most interesting relationships between them. In this work, we used the Gephi tool, which is a free software available at (https://gephi.org). To perform this step, the following tasks must be executed:

a) **Building a network:** With the Association Rules extracted in the previous step, we build a complex network. We use an undirected graph, since the direction of relationships is disregarded in this work, and the order of words is not important because we are not analyzing cause-effect relationships between the terms.

b) **Metric extraction:** Networks have a set of metrics that can describe different behavior and characteristics, such as node size, hubs, average connectivity, shortest path, diameter, clustering coefficient, proximity centrality, and the centrality of intermediation, among others. In this work, since we are interested in the centrality of nodes (terms), we choose the hubs, degree and triangle metrics.

c) **Ranking of keywords:** The metrics obtained in the previous step can help the identification of keywords that stand out in the network. Through a list in descending order of corresponding values, we obtain a ranking of the most representative keywords (terms) in relation to each of the measures of interest.

d) **Extraction of relationships:** By inspecting the results from the ranking, it may be possible to infer from the network, for example, how the keywords are related, what are the terms connecting the same (if any), and how terms connect to them.

4 APPLICATION AND RESULTS

In this section, we use the methodology to extract the keywords from 18 texts in the KG domain of relevant authors, some indicated in the work of de Sá Freire et al. (2017) (Table 1). In their work, Sá Freire et al. performed a bibliometric analysis to understand the evolution of scientific production in the KG domain. After the exploratory phase, they performed a descriptive analysis of articles adhering to the theme.

Table 1 – Work related to KG.

Authors and publication date	Topics
Santiso (2001)	international co-operation, good governance
Gooderham, Minbaeva, & Pedersen (2011)	governance mechanisms, social capital, knowledge transfer
Mayer (2006)	spillovers, governance, information technology
Antonelli (2006)	business governance, information economics
Antonelli (2007)	technological knowledge
Antonelli, Amidei, & Fassio, (2014)	mechanisms of knowledge governance
Antonelli (2016)	Schumpeterian growth model, technological change
Antonelli & Fassio (2016)	globalization, knowledge-driven economy
Pemsel et al. (2013)	knowledge governance, project-based organizations
Michailova & Sidorova (2011)	group-based work, organizational learning, knowledge sharing
Fong & Chen (2011)	governance of learning mechanisms
Grandori (1997)	governance structures, coordination mechanisms, cognitive models
Grandori (2001)	hierarchy, knowledge-governance mechanisms, theory of the firm
Buuren (2009)	inclusive knowledge management, collaborative governance
Chen & Fong (2012)	performance heterogeneity, knowledge management
Nooteboom (2000)	learning, absorptive capacity, cognitive distance, governance
Cole (2011)	global governance of knowledge, patent offices
Pemsel & Müller (2012)	knowledge governance, project-based organizations

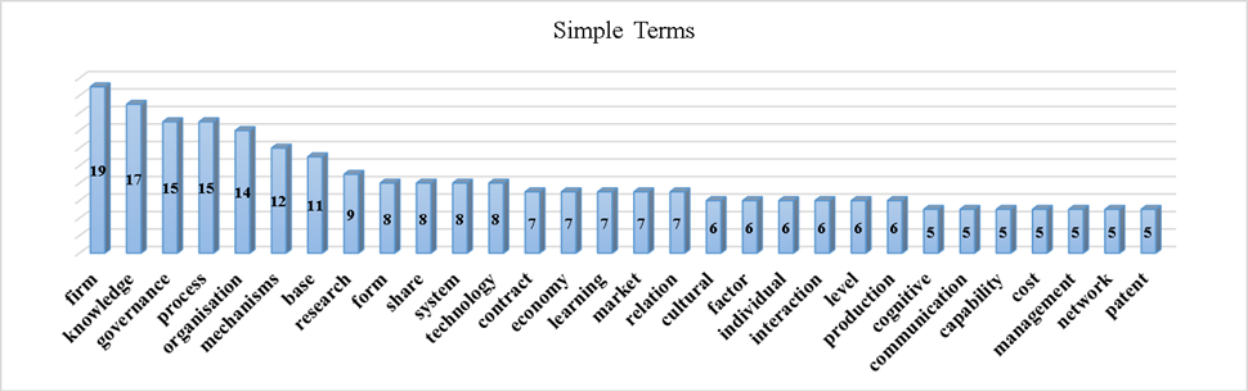
Source: extended from de Sá Freire et al. (2017).

Following the selection of texts, we inspected them to extract only the excerpts on the topic "Knowledge Governance". In most cases, this theme was contained in the section devoted to the state of the art. Each text was then saved as a new document to be processed separately to maintain the representativeness of a term within the textual context in which it was used.

We moved to the data preprocessing and the Association Rules extraction phase with the support of the FEATuRE tool (Rossi & Rezende, 2011). We opted for removing stop words to decrease word dimensionality. The documents were then mapped into transactions. We used mapping by selecting the sliding windows option with size five and step one. We chose to use sliding windows since this type of mapping can generate more transactions than sentence or paragraph mapping. Some examples of transactions are {paper}, {paper conceptualizes}, {paper conceptualizes defines} and {paper conceptualizes defines knowledge}.

From these transactions, we extract the rules. Some examples of rules are *knowledge* <- (25.6, 25.6); *knowledge* <- *learning* (1.3, 45.2); *learning* <- *knowledge* (1.3, 5.2), where the first value in parenthesis is the support and the second the confidence. From the rules, we may obtain the itemsets (e.g. *knowledge; based_knowledge; knowledge_processes*). At the end of this process, from the set of frequent itemsets, we obtain a list of all the simple (i.e. one-word) keywords that are contextually more frequent (Figure 1).

Figure 1 - Keywords composed by simple terms with their contextual frequency (25 shown).

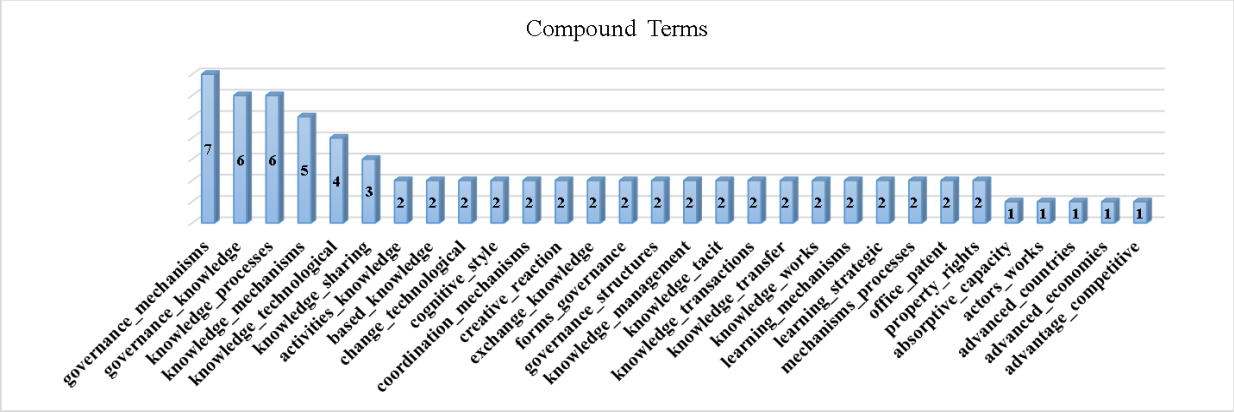


Source: the authors (2019).

From Figure 1, it is observed that some terms were already expected in the list (e.g. *firm*, *knowledge*, *governance*, etc.). However, terms (keywords) such as *research, communication,* and *production*, were unexpected. The uncovering of these keywords may be regarded as a partial result to be explored in further analysis. In addition, the compound, i.e. two or three-word keywords representing the domain, were also obtained through the set of frequent itemsets (Figure 2).

These terms are more contextualized than the former ones in Figure 1, as they reveal words that frequently appear together, such as *coordination mechanisms, cognitive style,* and *knowledge transactions*. Thus, this analysis automatically shows what are the terms that mostly support and compose the KG domain. Among those, we may also identify the five most predominant terms that define the domain, e.g. *governance mechanisms, governance knowledge, knowledge processes, knowledge mechanisms,* and *knowledge technological.*

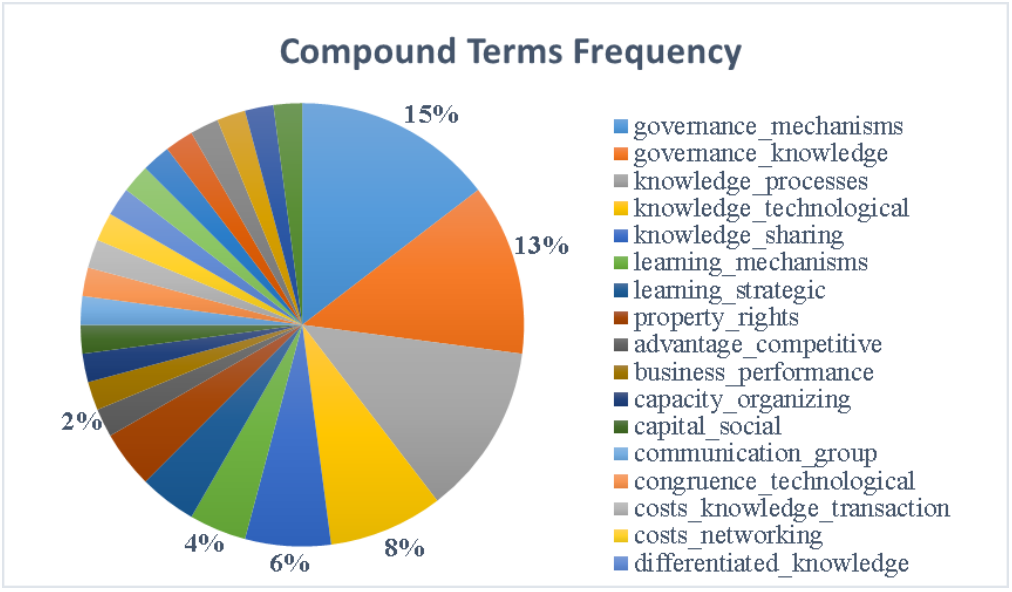
Figure 2 – Compound keywords (terms) and contextual frequency (25 shown).



Source: the authors (2019).

Figure 3 shows a pie chart with several compound terms and their frequency in the KG domain. It highlights how KM and KG are related, e.g. the terms *knowledge processes*, and *knowledge mechanisms* are also the ones that are the most predominant in KM.

Figure 3 – Compound keywords with their actual frequency for the KG domain.



Source: the authors (2019).

It is also important to stress out how *knowledge sharing* and *learning mechanisms* are fundamental to this domain. The first two terms are not new since they are directly related to the domain under analysis. However, the following terms (i.e. *knowledge-processes, knowledge-*

technological, knowledge-sharing) are related to KM area. *Learning-mechanisms* and *learning-strategic* are related to the field of organizational learning.

The graph shows that KM and organizational learning are two large areas that support KG. For a domain expert, this is an expected result which may be regarded as a form of validation of the method. For the general practitioner in the field without domain expertise, this may be deemed as a new information/knowledge. This fact highlights the possibility of using the method to explore knowledge domains that are still evolving and for which there is no common understanding of its foundations.

Moving to Step three of the methodology, i.e. creation of complex network from the compound terms, from the analysis of the Association Rules obtained in the previous step, we modeled a complex network with the aid of the Gephi tool (https://gephi.org/). Each word of a compound keyword is mapped onto a network node. For example, for the term *governance_mechanism*, the term *governance* yields a node and *mechanism* another one, and they are both connect by a link.

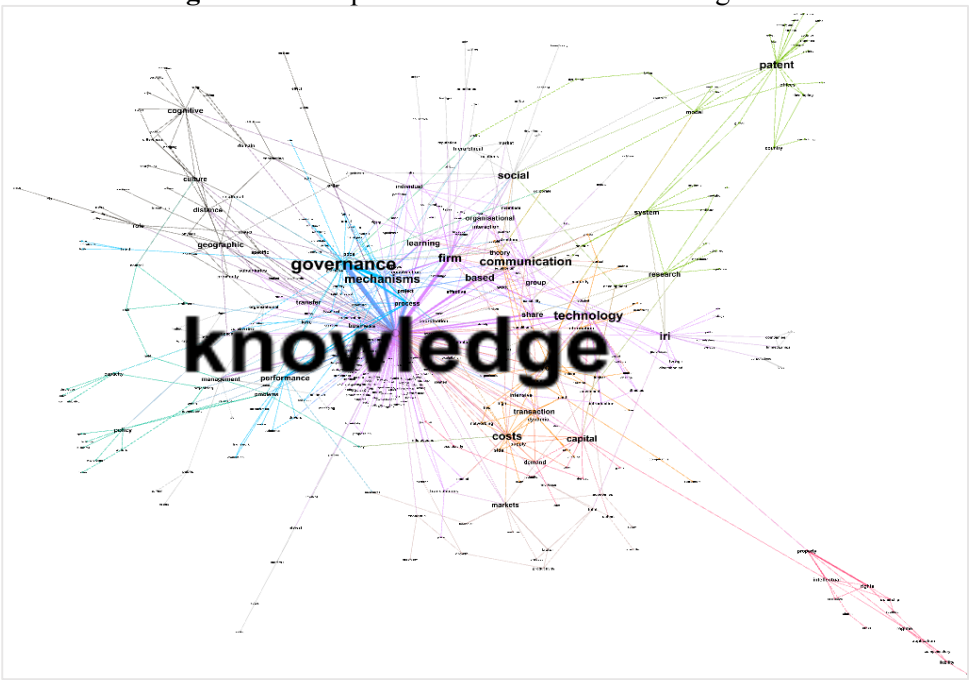
The general context of the network presented a non-directed graph containing 209 nodes and 438 edges (Figure 4). For the visualization of the network, we opted for the application of the *Force Atlas* algorithm, due to its synthetic way of displaying the nodes and connections, thus facilitating the visualization of their organization pattern. *Force Atlas* is an algorithm that uses the force of network relations as the basis for zooming in or out of nodes.

Next, we extracted the complex network statistics and then moved onto the analysis of the network properties (metrics). For the development of this experiment, we used the following metrics: the average degree of a vertex (number of connections); hubs (vertices with the highest connection intensity) and triangle (sets of three vertices connected to each other, representing grouping). Table 2 shows the evaluation metrics.

It can be seen from Table 1 that the term *communication*, which was ranked among the last 6 terms in Figure 1, was promoted to the fourth and fifth position. This result uncovers and highlights the role of *communication* within the KG domain. Whereas this role may be understood within the field, we argue that its actual strength is not easy and readily disclosed without the methodology here presented. In Figure 5, which is a fragment of the full network, we can see how the term *communication* is the bridge that links the other fundamental terms of the domain. The

processing time of the tests performed was in the order of a few seconds in an Intel (R) Core (TM) i3-3217U CPU @ 1.80 GHz processor.

Figure 4 – Complex network based on node degree.



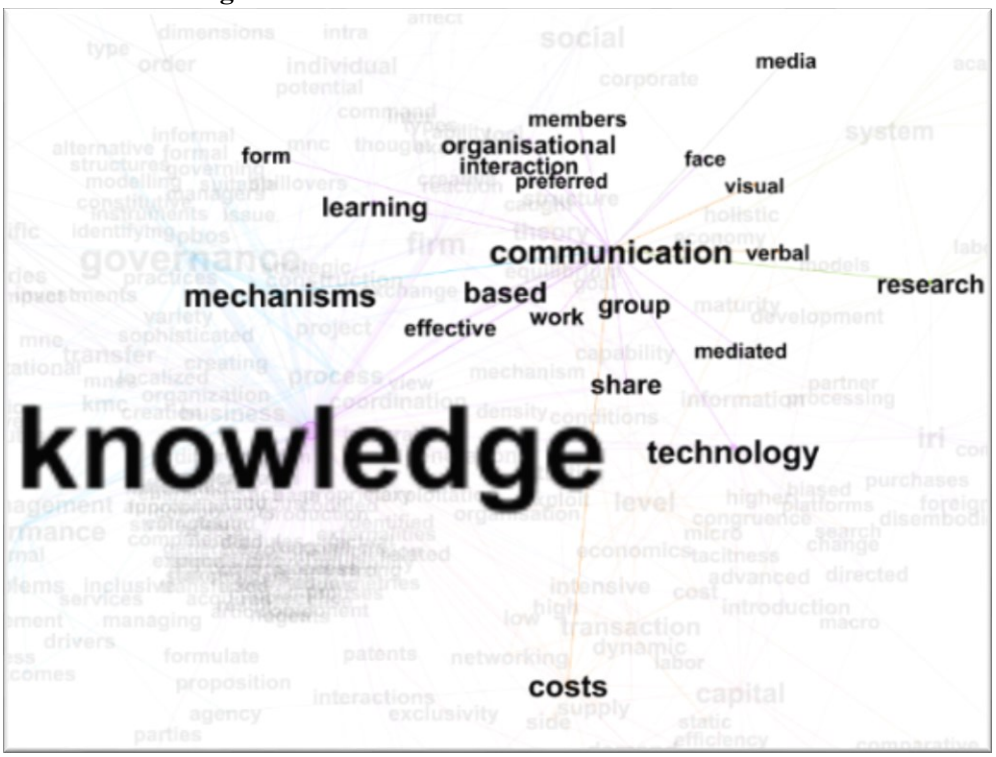
Source: the authors (2019).

Table 2 – Ranking of terms.

TERM	DEGREE	TERM	HUB	TERM	TRIANGLE
Knowledge	173	knowledge	0,6350	knowledge	160
Governance	38	governance	0,1413	governance	42
Social	24	mechanisms	0,1413	social	37
Mechanisms	23	**communication**	0,1238	mechanisms	35
Communication	23	costs	0,1136	**communication**	34
Capital	22	capital	0,1110	costs	28
Patent	22	based	0,1041	capital	28
Costs	19	learning	0,0990	baded	22
Based	17	sharing	0,0898	learning	22
Performance	14	group	0,0870	group	19
Technological	14	business	0,0849	sharing	18
Learning	13	organisational	0,0846	organisational	18
Geographic	13	transaction	0,0838	transaction	18
Firm	12	transfer	0,0790	geographic	18
Cognitive	12	geographic	0,0783	business	13
Levels	11	processes	0,0781	transfer	13
System	11	interaction	0,0771	firms	13
Transaction	11	firms	0,0769	distance	13
Research	11	technological	0,0751	work	12
Markets	10	work	0,0746	processes	11

Source: the authors (2019).

Figure 5 – *Communication* as the central node.



Source: the authors (2019).

5 CONCLUSION

Technological approaches such as the use of artificial intelligence techniques are a knowledge governance mechanism that can facilitate knowledge sharing processes. This kind of activity can stimulate internal creativity and innovation. In this context, the exploitation of knowledge in public academic research documents mediated by technological mechanisms can be the basis for discoveries at a reduced cost, thus supporting Knowledge Management and Knowledge Governance systems in organizations. Through this technology, the explicit knowledge available can produce new knowledge that favors the competitive advantages of the market.

One key tool for Knowledge Management is the ability to automatically perform text analysis in a collection of documents. In this work, the main goal was essentially to turn text into data for analysis, via the application of Text Mining techniques and analytical methods. Specifically, we have considered a set of academic and public documents related to the field of Knowledge Governance. From these documents and using the presented methodology, we have

identified a set of keywords that characterize the Knowledge Governance domain. From these keywords, we have found some of the most predominant simple and compounds terms.

We have found that the term "*communication"* was one of the most related terms (with other keywords) in Knowledge Governance. This analysis would not be possible to be carried out by a human reader in such a short frame of time. The proposed approach allows the domain expert to focus on the analysis of the terms found in order to extract new information and knowledge.

As future work, we aim at analyzing the computational performance of the approach by escalating the number of documents to a large set of documents with the support of additional automation tools. We also aim at the implementation of the semantic analysis of the terms and the extension of the tool with visualization aids.

Based on the experience from this project, the authors estimate that the cost of investing in this approach is feasible in view of training, qualified staff, and IT infrastructure. The processing time of the tests performed was in the order of a few seconds, which is also an indicator of feasibility in terms of hardware resources. The bottleneck was the gathering of the documents. However, work in progress is already addressing this issue, and preliminary results are encouraging.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

Antonelli, C. (2006). The business governance of localized knowledge: an information economics approach for the economics of knowledge. *Industry and Innovation*, *13*(3), 227-261.

Antonelli, C. (2007). Technological knowledge as an essential facility. *Journal of Evolutionary Economics*, *17*(4), 451-471.

Antonelli, C., Amidei, F. B., & Fassio, C. (2014). The mechanisms of knowledge governance: State owned enterprises and Italian economic growth, 1950–1994. *Structural Change and Economic Dynamics*, *31*, 43-63.

Antonelli, C. (2016). A Schumpeterian growth model: wealth and directed technological change. *The Journal of Technology Transfer*, *41*(3), 395-406.

Antonelli, C., & Fassio, C. (2016). Globalization and the knowledge-driven economy. *Economic Development Quarterly*, *30*(1), 3-14.

Aranha, C., & Passos, E. (2006). A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, *5*(2).

Ashok, M., Narula, R., & Martinez-Noya, A. (2016). How do collaboration and investments in knowledge management affect process innovation in services? *Journal of Knowledge Management*, *20*(5), 1004-1024.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences*, *39*(1), 1-20.

Becerra-Fernandez, I., & Sabherwal, R. (2014). *Knowledge management: Systems and processes*. Routledge.

Berlingerio, M., Bonchi, F., Bringmann, B., & Gionis, A. (2009, September). Mining graph evolution rules. In joint *European conference on machine learning and knowledge discovery in databases* (pp. 115-130). Springer, Berlin, Heidelberg.

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.

Bhardwaj, P., & Khosla, P. (2017). Review of text mining techniques. *IITM Journal of Management and IT*, *8*(1), 27-31.

Bouakkaz, M., Ouinten, Y., Loudcher, S., & Strekalova, Y. (2017). Textual aggregation approaches in OLAP context: A survey. *International Journal of Information Management*, *37*(6), 684-692.

Bouakkaz, M., Ouinten, Y., Loudcher, S., & Fournier-Viger, P. (2018). Efficiently mining frequent itemsets applied for textual aggregation. *Applied Intelligence*, *48*(4), 1013-1019.

Boudin, F. (2013, October). A comparison of centrality measures for graph-based keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 834-838).

Buuren, A. V. (2009). Knowledge for governance, governance of knowledge: Inclusive knowledge management in collaborative governance processes. *International public management journal*, *12*(2), 208-235.

Chang, J. Y., & Kim, I. M. (2013). Analysis and evaluation of current graph-based text mining researches. *Advanced Science and Technology Letters*, *42*, 100-103.

Chen, H. (2001). *Knowledge management systems: a text mining perspective*. Knowledge Computing Corporation.

Chen, L., & Fong, P. S. (2012). Revealing performance heterogeneity through knowledge management maturity evaluation: A capability-based approach. *Expert Systems with Applications*, *39*(18), 13523-13539.

Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, *34*(2), 272-284.

Cleveland, G. (1998). Digital libraries: definitions, issues and challenges.

Cody, W. F., Kreulen, J. T., Krishna, V., & Spangler, W. S. (2002). The integration of business intelligence and knowledge management. *IBM systems journal*, *41*(4), 697-713.

Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., & Smalheiser, N. R. (2010, November). Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In *Proceedings of the 1st ACM international Health Informatics Symposium* (pp. 376-380). ACM.

Cole, J. H. (2011). The global governance of knowledge: patent offices and their clients, Prometheus, 29:1, 51-54, DOI: 10.1080/08109028.2011.567852.

Daugherty, P. R., & Wilson, H. J. (2018). *Human+ machine: reimagining work in the age of AI*. Harvard Business Press.

de Sá Freire, P., Dandolini, G. A., de Souza, J. A., Caetano Silva, T., & Moreira Couto, R. (2017). Governança do Conhecimento (GovC): o estado da arte sobre o termo. *Biblios*, (69), 21-40.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63-71.

Fan, W., Wang, X., Wu, Y., & Xu, J. (2015). Association rules with graph patterns. *Proceedings of the VLDB Endowment*, *8*(12), 1502-1513.

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

Ferraris, A., Santoro, G., & Dezi, L. (2017). How MNC's subsidiaries may improve their innovative performance? The role of external sources and knowledge management capabilities. *Journal of Knowledge Management*, *21*(3), 540-552.

Fong, P. S., & Chen, L. (2011). Governance of learning mechanisms: Evidence from construction firms. *Journal of construction engineering and management*, *138*(9), 1053-1064.

Foss, N. J. (2007). The emerging knowledge governance approach: Challenges and characteristics. *Organization*, *14*(1), 29-52.

Foss, N. J., Husted, K., & Michailova, S. (2010). Governing knowledge sharing in organizations: Levels of analysis, governance mechanisms, and research directions. *Journal of Management studies*, *47*(3), 455-482.

Gooderham, P., Minbaeva, D. B., & Pedersen, T. (2011). Governance mechanisms for the promotion of social capital for knowledge transfer in multinational corporations. *Journal of Management Studies*, *48*(1), 123-150.

Grandori, A. (1997). Governance structures, coordination mechanisms and cognitive models. *Journal of Management & Governance*, *1*(1), 29-47.

Grandori, A. (2001). Neither hierarchy nor identity: knowledge-governance mechanisms and the theory of the firm. *Journal of management and Governance*, *5*(3-4), 381-399.

Grant, R. M. (1996). Prospering in dynamically competitive environments: Organizational capability as knowledge integration. *Organization science*, *7*(4), 375-387.

Hashimi, H., Hafez, A., & Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, *51*, 729-733.

Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. *Journal of Business Economics*, *87*(3), 317-335

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: human-AI symbiosis in organizational decision-making. *Business Horizons*, *61*(4), 577-586.

Jiang, H., Horner, H. T., Pepper, T. M., Blanco, M., Campbell, M., & Jane, J. L. (2010). Formation of elongated starch granules in high-amylose maize. *Carbohydrate Polymers*, *80*(2), 533-538.

Jones, K. S. (1999). Information retrieval and artificial intelligence. *Artificial Intelligence*, *114*(1-2), 257-281.

Krapivin, M., Autayeu, A., Marchese, M., Blanzieri, E., & Segata, N. (2010, June). Keyphrases extraction from scientific documents: improving machine learning approaches with natural language processing. In *International Conference on Asian Digital Libraries* (pp. 102-111). Springer, Berlin, Heidelberg.

Lahiri, S., Choudhury, S. R., & Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. *arXiv preprint arXiv:1401.6571*.

Liébana-Cabanillas, F., Marinković, V., & Kalinić, Z. (2017). A SEM-neural network approach for predicting antecedents of m-commerce acceptance. *International Journal of Information Management*, *37*(2), 14-24.

Liu, Y., Chan, C., Zhao, C., & Liu, C. (2018). Unpacking knowledge management practices in China: Do institution, national and organizational culture matter? Published online. *Journal of Knowledge Management.*

Mao, H., Liu, S., Zhang, J., & Deng, Z. (2016). Information technology resource, knowledge management capability, and competitive advantage: The moderating role of resource commitment. *International Journal of Information Management*, *36*(6), 1062-1074.

Mayer, K. J. (2006). Spillovers and governance: An analysis of knowledge and reputational spillovers in information technology. *Academy of Management Journal*, *49*(1), 69-84.

Marwick, A. D. (2001). Knowledge management technology. *IBM systems journal*, *40*(4), 814-830.

Michailova, S., & Foss, N. J. (2009). Knowledge governance: themes and questions. *Knowledge governance: Processes and perspectives*, 1-24.

Michailova, S., & Sidorova, E. (2011). From group-based work to organizational learning: the role of communication forms and knowledge sharing. *Knowledge Management Research & Practice*, *9*(1), 73-83.

Monks, R. A. G., & Minow, N. (1995). Corporate governance on equity ownership and corporate value. *Journal of financial Economics*, *20*, 293-315.

Nahapiet, J. and Ghoshal, S. (1998). 'Social capital, intellectual capital and the organizational advantage'. *Academy of Management Review*, 23, 242–66.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521-543.

Nooteboom, B. (2000). Learning by interaction: absorptive capacity, cognitive distance and governance. *Journal of management and governance*, *4*(1-2), 69-92.

Pemsel, S., & Müller, R. (2012). The governance of knowledge in project-based organizations. *International Journal of Project Management*, *30*(8), 865-876.

Pemsel, S., Wiewiora, A., Müller, R., Aubry, M., & Brown, K. (2013). Knowledge Governance in Project-Based Organizations: Theoretical and Empirical Approaches. In *IRNOP-International Research Network on Organizing by Projects 11*.

Pemsel, S., Wiewiora, A., Müller, R., Aubry, M., & Brown, K. (2014). A conceptualization of knowledge governance in project-based organizations. *International Journal of Project Management*, *32*(8), 1411-1422.

Pietersen, W. (2010). Strategic learning: How to be smarter than your competition and turnkey insights into competitive advantage. *John Wiley & Sons*.

Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management*, *38*(1), 201-216.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., & Paula, M. D. (2003). Mineração de dados. *Sistemas inteligentes: fundamentos e aplicações*, *1*, 307-335.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, *1*, 1-20.

Rossi, R. G., Rezende, S. O.(2011). Generating features from textual documents through association rules. *Anais do Encontro Nacional de Inteligência Artificial. SBC.*

Santiso, C. (2001). International co-operation for democracy and good governance: moving towards a second generation?. *The European Journal of Development Research*, *13*(1), 154-180.

Santoro, G., Vrontis, D., Thrassou, A., & Dezi, L. (2018). The Internet of Things: Building a knowledge management system for open innovation and knowledge management capacity. *Technological Forecasting and Social Change*, *136*, 347-354.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Sonawane, S. S., & Kulkarni, P. A. (2014). Graph based representation and analysis of text document: A survey of techniques. *International Journal of Computer Applications*, *96*(19).

Spender, J. C. (1996). Making knowledge the basis of a dynamic theory of the firm. *Strategic management journal*, *17*(S2), 45-62.

Stenmark, D. (2002, January). Information vs. knowledge: The role of intranets in knowledge management. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 928-937). IEEE.

Tan, P. N., Steinbach, M., Kumar, V., & Karpatne, A. (2019). *Introduction to Data Mining, Global Edition*. Pearson Education Limited.

Tsang, E. W. K. (2000). 'Transaction cost and resource-based explanations of joint ventures: a comparison and synthesis'. *Organization Studies*, 21, 215–42.

Wu, J. L., & Agogino, A. M. (2004, January). Automating keyphrase extraction with multi-objective genetic algorithms. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (pp. 8-pp). IEEE.

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., & Boccaletti, S. (2016). Combining complex networks and data mining: why and how. *Physics Reports*, *635*, 1-44.