A MapReduce-Based Method for Achieving Active Technological Surveillance in Big Data Environments

Daniel San Martin Pascal Filho, Douglas Dyllon Jeronimo de Macedo, and Moisés Lima Dutra(✉)

PGCIN, Federal University of Santa Catarina, Florianopolis, Brazil
{douglas.macedo,moises.dutra}@ufsc.br

Abstract. Technological Surveillance systems stand out as a structured way to assist organizations in monitoring their internal and external technological environments, in order to anticipate changes. However, since the volume of digital data available keeps growing, it becomes increasingly complex to keep this type of system running without proper automation. This paper proposes an automated MapReduce-based method for technological Surveillance in Big Data scenarios. A prototype was developed to monitor key technologies in specialized portals in the Furniture and Wood sector, in order to illustrate the proposed method. The proposal was evaluated by industry experts, and the preliminary results obtained are very promising.

Keywords: Technological surveillance · MapReduce · Big Data · Ontologies

1 Introduction

Throughout history, technological changes have created, transformed and destroyed markets. The monitoring of the main technological trends plays a key role in increasing the competitiveness of companies. However, the accelerated pace of technological development and the steady reduction of time between different innovation cycles increase complexity and the efforts required to maintain an up-to-date picture of the whole technological scenario [23].

In search of solutions, a wide variety of approaches to technological monitoring have been proposed. They range from interviews, through simple keyword-based Internet searches, to sophisticated text mining systems in search of hidden patterns [7]. Among those, Technological Surveillance (TS) has gained prominence. TS seeks to establish a monitoring process that ranges from collecting data to communicating insights about it to decision makers. That allows organizations to have an overview of the technological changes from different sources of information. Nevertheless, maintaining a functional and active monitoring system to evaluate real Big Data scenarios composed of patents, scientific papers,

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 254–271, 2021.
https://doi.org/10.1007/978-3-030-77417-2_19

books, and spreadsheets, to name a few, is still one of the biggest challenges faced by organizations.

In this sense, the objective of this work is to propose a method to identify the use of existing technologies in Big Data scenarios, especially in those environments densely populated by large volumes of data. In a nutshell, we intend to answer the following research question: how to automatically identify a set of technologies of interest whose use is increasing in popularity on Web portals?

The remainder of this work is organized as follows. In Sect. 2, a theoretical framework of reference is given, which depicts the concepts of technological surveillance, and also talks about information, domain, Big Data scenarios, and the MapReduce model. Section 3 presents the methodological procedures used to develop this work. In Sect. 4, the proposed method is described, along with the prototype developed to implement it and the experimental results achieved. Finally, Sect. 5 presents the final considerations and suggestions for future works.

2 Theoretical Framework

2.1 Technological Surveillance

Technological Surveillance (also known as Technological Vigilance or Technological Watch) arose from the need to monitor the internal and external environments of organizations in a structured and systematic way, in order to map possible changes in technological scenarios.

Table 1 presents the most relevant TS definitions in the context of this work. They converge in terms of the proposed/existing stages of the surveillance process, raging across collecting information, analyzing it, and communicating the results obtained to interested parties. Palop and Vicente [21] point out the key reasons that lead organizations to adopt a TS system: (i) the need to anticipate changes in order to avoid competitive disadvantage, especially in scenarios where technology can be a differential; (ii) cost reduction; (iii) the progress of the organization in relation to the market; (iv) the need to innovate; and (v) new possibilities to establish cooperation with other organizations. In summary, it can be stated that Technological Surveillance is a method used to monitor and evaluate different sources of information, mostly of which to be used as an input for patent registration processes. That is, it searches for evidences that could signal changes in the technological scenario in which an organization is inserted.

In the literature, proposals for automated technological monitoring are presented by using patent bases [17,29], such as the database of the USTPO (US Patent and Trademark Office) [9,18], scientific papers [1,11], and publications available on Web portals [26]. The techniques used by the authors are mainly focused on the extraction of terms from text documents, by means of calculating the TF-IDF (term frequency–inverse document frequency) and applying the LDA (Latent Dirichlet Allocation) algorithm for topic modeling [17,18,29]. Besides, several techniques for document clustering are applied, notably the k-means algorithm [9,18], in order to detect and recognize outstanding technologies.

Table 1. Definitions of technological surveillance.

Definition	Authors
Method of collecting, analyzing, disseminating, communicating and using information for decision making, including competitive intelligence or similar terms	Palop and Vicente (1999) [21]
A structured system to coordinate the activities of information retrieval, analysis and dissemination, both inside and outside the organization, according to an organizational plan and strategy	Salgado Batista et al. (2003) [4]
Technological surveillance goes through the stages of diagnosis, research and capture of information, analysis of information, valuation of relevant information, information dissemination and communication, by offering guidance to decision making	OVTT [20]
A technology surveillance model consists of a set of processes: identification of needs, a definition of sources and means of access to information; search, processing and validation of information; and valuing of information, results, measurement and improvements	AENOR [3]
A systematic process that aims to identify, organize and correlate the results of technological prospecting in order to make them useful to the organization's strategies	ABNT [2]

However, these proposed methods limit their field of monitoring, since most organizations are immersed in Big Data scenarios, in which they need to store and monitor a large volume of data in a variety of formats, such as text documents, database content, audio, video, spreadsheets, patents, clicks, physical-device data, internal systems data, Web portal content, event records, news-website content, social-media content, scientific publications, among others.

2.2 The Information and Its Sources

From the point of view of Information Science (IS), according to Buckland's view [5], information in the process of Technology Surveillance can be understood as knowledge and the analyzed documents as "things" that can reduce uncertainties. Therefore, the ability to evaluate the quality of the documents and correctly delimit their sources is fundamental to this process, being one of the factors that most impact surveillance results, according to León et al. (2006) [16].

According to quality criteria presented by Muñoz et al. (2006) [14], sources of information can be divided into formal, such as articles or books, and informal, such as talks or visits to fairs. A TS system should work with formal sources of information. As far as processing is concerned, sources of information can be classified as "electronically available" and "not electronically available".

The quality of information sources can be analyzed by means of applying specific criteria and indicators. In her book "Information Sources on the Internet", Tomaél (2008) [28] lists as possible criteria to be used: information architecture and its intrinsic aspects (content *versus* user need), the credibility of the information source (contextual handling aspects), and information representation (conciseness and consistency) and its sharing aspects.

2.3 Domain

The concept of domain was introduced in the IS context by Birger Hjørland (1995) [12]. Before that, it was already widely used in Computer Science, in which was disseminated in the mid-1980s [19]. A domain is a group that shares an ontology, undertakes common research or work, and also engages in discourse or communication, formally or informally [27].

The analytical domain paradigm of Information Science states that the best way to understand the concept of information in IS is studying the domains of knowledge as communities of thought or discourse, which are part of the division of labor of society [12]. According to the domain analysis, if we wanted, for example, to develop a software application for Brazilian Geography analysis, we should not focus on certain users, but instead call a geographer specialized in Brazilian Geography to help us. Going against this premise, this research used information produced by experts from a particular industry sector to model ontologies containing the key technologies and critical resources to their area.

2.4 Big Data

The information of interest in Technological Surveillance systems are usually collected from different sources and possess different formats. TS systems need to deal with thousands of unstructured digital documents, as well as huge volumes of data that, due to the high level of information digitization, keep growing. In this work, we consider these configurations as part of Big Data scenarios.

Among the several existing Big Data definitions found in the literature, two stand out: the definitions 3 "Vs" and 5 "Vs". The 3 "Vs" definition was adopted by authors such as [13] and [15], which characterize it as being: variety, and velocity. The definition of the 5 "Vs" adds two new features to them: veracity and value [30]. There are still other definitions composed by more "Vs", and even other letters. Actually, there is no consensus at all about the concept of Big Data. In Technological Surveillance scenarios, data is only useful if it has value and can be checked for accuracy. Consequently, in this work we adopted the definition of 5 Vs to define Big Data scenarios.

Unlike traditional database management systems (DBMSs), such as MySQL or Oracle, which were built to work with structured data in the form of relational tables, Big Data data often varies in structure. Therefore, operations as text mining are often common and required to process or retrieve information. Partitioning data across multiple computers or nodes may also be required to

manage the large volume of data or allow it to be processed and analyzed in an acceptable time frame.

2.5 MapReduce Processing in Big Data Scenarios

MapReduce was conceived in Google as a solution to optimize its Web search engine in 2003, based on the paradigm of functional computing and programming languages such as LISP. Between 1999 and 2004, Google engineers developed dozens of algorithms to process data in order to generate more data. In 2008, the model was already scaled, processing 20 petabytes daily [6]. According to Dean and Ghemawat [6], the programming model MapReduce is an associated implementation useful for a wide variety of real world tasks and was designed to handle large datasets. Its structure enables parallel processing in large collections of data through computer clusters. Currently, it is the main programming method used to work in Big Data scenarios.

Through it, it is possible to elaborate simple algorithms capable of being executed both in a personal computer and in distributed environments composed of thousands of computers, allowing the construction of scalable solutions. Otherwise, the processing would be sequential, possibly presenting performance problems in environments with large volumes of data. Conceptually, the Map function receives a collection of input data and applies a function on them, generating a new collection. The Reduce function receives a collection of input data and applies a function on them, providing a reduction in the size of the collection.

Figure 1 presents the application of the MapReduce model for counting words. The input sentences are splitted into words, for which the value 1 is assigned during the Map step. In the Reduce step, similar words have their values summed and assigned. As a result, a list is generated with the words and the total number of their occurrences in the input text.

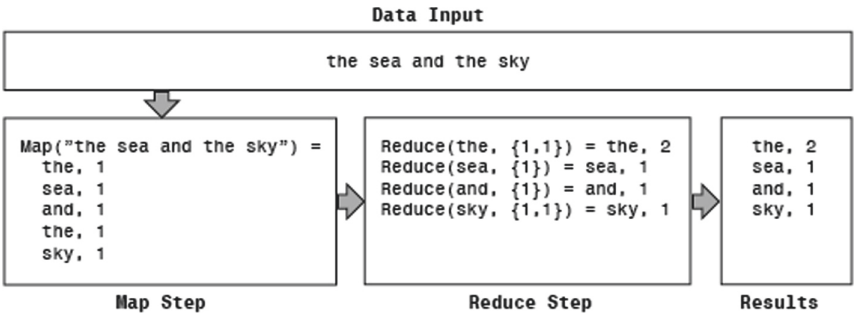

Fig. 1. MapReduce example.

3 Methodological Procedures

This work is an applied research, as it aims to generate knowledge for the application of methods aimed at solving a specific problem involving local interests. Regarding the objectives, among the classifications proposed by Gil [10], this work conducts an explanatory research, which seeks to understand the causes and effects of the phenomena under study, in a structured, quantitative and lateral way in experimental methods. To overcome the limits of the quantitative research [8], this work also makes use of a qualitative approach. When validating the proposal with specialists, a qualitative evaluation of the result is necessary. Quantitative approaches are also used in the data analysis process, during the experiments.

This work uses experimental and bibliographic procedures, too. The bibliographical research seeks to explain a problem from theoretical references published in documents, such as books, periodicals, scientific papers, without the elaboration of hypotheses. This approach is important mainly in the theoretical foundation, by bringing knowledge to the authors and structuring the basis for the construction of the remaining knowledge of this work. According to Gil [10], the experimental research consists in determining a study object, selecting the variables that would be able to influence it, and defining the ways of controlling and monitoring the effects produced by the variables on the object.

The elaboration of this research takes place through the definition of a research problem and a research question, in order to establish the paper's goals. In addition, in order to support the early research process, topics like technology surveillance, information, domain, Big Data, and MapReduce are studied. Besides, an analysis of related works is carried out to compose the necessary theoretical basis of the work, which is used to objectively interpret and analyze the subject, as well as produce conclusions based on proven theories.

During the analysis of the related works, opportunities for improvement are identified. Subsequently, they are included in this work's main proposal. To validate it, a functional prototype is built, by using MongoDB technology as the data repository, Protegé as the software tool for ontology modeling, Python as the programming language for capturing, parsing and analyzing data, and Javascript as the programming language for implementing the MapReduce model. To validate this prototype, we present the results obtained from specialists who pointed out the correct alignment of the results in relation to the market.

4 The Proposed Method

Table 2 summarizes the proposed method for Technological Surveillance in Big Data scenarios. It aims to automatically identify key technologies and assess their popularity in large volumes of documents collected from formal sources of digital information.

It is based on the TS cycle proposed by Sánchez Torres and Palop Marro (2002) [22]. The main steps of the proposed method are: (1) **Planning**. Identification of the needs and objectives of the organization; (2) (3) **Collection**

and **Organization**. Gathering and filtering of the data, in order to select it and eliminate what is not necessary; (4) **Intelligence**. Qualification of information and its alignment with the organization's strategies; and (5) **Communication**. Dissemination of the results within the organization for the purpose of decision making.

The proposed method (Table 2) delimits a series of activities required to enable the automation process. The **Planning** step requires human intervention, which involves experts, decision makers, and anyone else interested in the outcome of the process, as it must be aligned with the strategic objectives of the organization. At this stage, a domain analysis is applied, by means of delineating the domains of interest and the technologies that should be monitored. The sources of digital information, such as Web portals, are also evaluated by using criteria indicated by Tomaél (2008) [28]. Finally, the technological knowledge of the domain is modeled in the form of ontologies, used to represent the terms to be located in the collected texts. For example, if there is interest in monitoring technologies related to energy production, the Energy domain could be modeled, by creating classes such as Renewable Energy and Non-renewable Energy. The Renewable Energy class, in turn, could aggregate subclasses like Solar Energy, Wind Energy, or Biomass Energy, which would represent the searched terms. This approach is not only useful because it maintains the hierarchy of concepts, but also because it facilitates the grouping of terms by levels of abstraction, i.e., the superclasses.

Table 2. The proposed method for technological surveillance.

Step	Main Features
1. Planning	Domain analysis. Ontology modeling. Assessment of digital sources of information
2. Collection	Web Crawlers. Web Scrapers. Query APIs. Gathering of digital data
3. Organization	Data wrangling. Storage in non-relational databases
4. Intelligence	MapReduce application. Identification of key technologies. Technology counting. Construction of time series. Creation of charts
5. Communication	Report generation. E-mail sending. Business Intelligence tools. Website feed

The **Collection** step is responsible for automatically collecting publications from predefined sources and make them available for the next step. It demands the construction of web robots, known as Web Crawlers [25]. Web crawlers gather data through a process known as Web Scraping. Moreover, publication data can also be collected either through APIs (Application Programming Interfaces) provided by the content portals, or by the directly collecting of digital documents from specialized repositories. All the collected data are stored in its raw form.

During step 3, **Organization**, a data wrangling process is performed. This process consists of extracting content from the collected publications, such as titles, dates, entities, among others. Each piece of extracted content is converted to a structured data type, like text, date, time, or numeric, in order to allow its further manipulation. Subsequently, they are organized and indexed in appropriate databases. In addition, the duplicated documents are removed in this step.

The texts extracted from the collected documents still represent a challenge for the analysis process, mainly due to its high dimensionality. Much of the information contained in the publications are useless. Some words are not relevant to the TS process. Therefore, it is essential to identify and extract useful data from the texts. Figure 2 shows an example of text reduction during the content extraction process. On the left is the original document collected. On the right side are the data stored in the database, after the data wrangling process. In the end, the reduced data refer to technologies identified with the analyzed scenario.

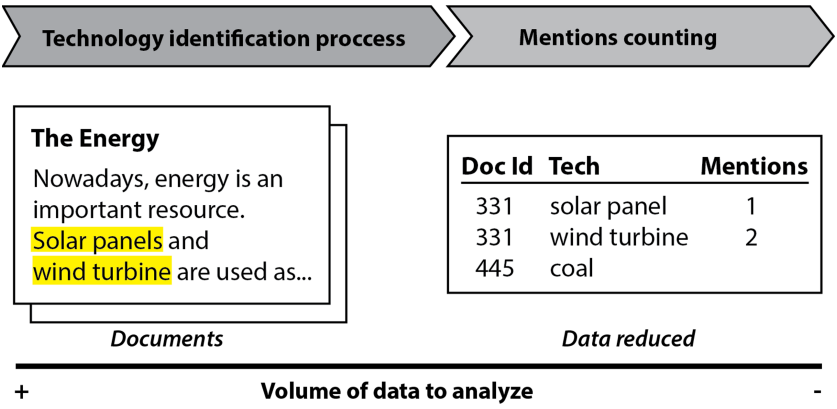

Fig. 2. Dimensionality reduction.

In Big Data contexts, moving large amounts of data for processing is quite computationally expensive. MapReduce functions possess the advantage that they can be executed in parallel on each cluster node where the data is. In this sense, the **Intelligence** step uses a MapReduce algorithm to identify key technologies, in order to enable the TS system to cope with Big Data. The algorithm generates a data subset with less dimensionality, which contains the technologies found in the source publications.

Once the volume of technologies has been quantified, it is possible to build time series that correlate the mentions of the technologies identified with their publication dates. Next, graphs that summarize the calculations performed can be generated to facilitate the analysis of the specialists.

Finally, the communication stage aims to satisfy users' information needs through the generation or automatic update of by-products of technological surveillance. In the previous steps, data were collected, analyzed and transformed, making them more understandable. Now, the interpretation of the final users must be facilitated by the enrichment and formatting of this new information, through the generation of reports containing the produced charts. Other examples of TS by-products are: (i) sending alerts by e-mail, (ii) updates to a specific BI system, (iii) and updates to a certain section of a website dedicated to the mining subject.

The proposed method defines a clear sequence of steps to create an automated system, which is designed to monitor a set of technologies of interest in electronically available information sources. This monitoring also includes checking how popular these technologies are on the Web. In the next section, we present a prototype that implements the proposed method and runs it on a given set of information sources.

4.1 Prototype

The purpose of the prototype is to validate the proposed method by implementing the proposed activities. The Furniture and Wood sector was chosen as a scenario because it is one of the key sectors in the economy of the Brazilian state of Santa Catarina. This sector is also included in the Industrial Development Program of Santa Catarina (PDIC 2022) [24], which is developed by the Federation of Industries of the State of Santa Catarina (FIESC). The architecture and main steps of the prototype are presented in Fig. 3.

The main technologies used for this prototype are Python language (along with its libraries) and MongoDB database. Python language ecosystem offers a set of libraries to work with data processing and visualization. MongoDB is a multi-platform open source database designed for documents classified as NoSQL. In it, documents are stored in JSON-style format. MongoDB was chosen because of its ability to scale horizontally, by adding new instances in other computers. Moreover, it possesses orientation to documents, since it allows the modeling of the various data collected during the TS process, such as scientific publications, patents, documents come from trade show and events, etc. MongoDB makes the representation of those documents closer to the analysts' reality, i.e. flexible for analysis. Besides, it provides the possibility of data processing through Map and Reduce functions.

Planning. In the Planning stage, a domain analysis was carried out on the Furniture and Wood sector, present in the PDIC 2022 notebook. Moreover, a discussion with FIESC specialists was held. Subsequently, it was modeled a domain ontology containing the key technologies listed for the this economy sector. Each technology was connected to a superclass that is used to group similar concepts. The ontology modeling was done in the Protegé tool, and the output was an OWL (Web Ontology Language) file. The selection of information sources was

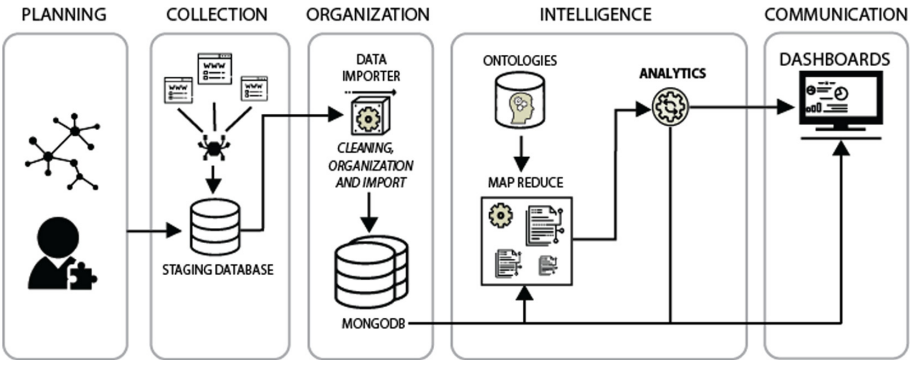

Fig. 3. Prototype architecture.

based on interviews with FIESC specialists, and also by reusing those already monitored by the FIESC Observatory, such as the Web portals: For Mobile[1], Furniture News[2], Mega Movers[3], Wood Business[4], and Woodworking Network[5].

Collection. A set of web crawlers was developed in the Python language to collect publications from the selected web portals. The main Python library used was BeautifulSoup[6], which offers specialized functions to access web pages and capture the desired content. Web crawlers work in two steps. First, they browse the websites where the publications are listed, collect their links, and store them. After, they visit each URL to scan the documents. Their titles, texts, and publication dates are collected and stored in a specific collection in the MongoDB database, as shown below.

Organization. The content extracted from the URLs visited are stored as documents in a collection called "crawled_news_urls". These documents contain metadata, e.g. an indicator of whether the content of a given document was extracted or not and its URL source. The content captured in the second stage of the Collection stage constitute another type of document, and are stored in a collection called "publications". Code snippets 1.1 and 1.2 below show examples of each type of document.

[1] http://www.formobile.com.br.
[2] http://www.furniturenews.net.
[3] http://www.megamoveleiros.com.br.
[4] http://www.woodbusiness.ca.
[5] http://www.woodworkingnetwork.com.
[6] Python package used to analyze documents in HTML and XML formats.

```
{
   "_id" : ObjectId("5b4ff6202e77ca12b95c98c8"),
   portal_name" : "Wood Business",
   "domain" : "FurnitureAndWood",
   "portal_url" : "https://www.woodbusiness.ca/industry-news/news/Page-8",
   "news_url" : "http://www.woodbusiness.ca/industry-news/news/the-new-bioeconomy-adding-value-to-
      biomass-4655",
   "extracted" : 1
}
```

Code Snippet 1.1. Sample document from the "crawled_news_urls" collection.

```
{
    "_id" : ObjectId("5b4ff9272e77ca132655b833"),
    "portal_name" : "Wood Business",
    "domain" : "FurnitureAndWood",
    "portal_url" : "https://www.woodbusiness.ca/industry-news/news/Page-2",
    "news_url" : "http://www.woodbusiness.ca/industry-news/news/canfor-appoints-dianne-watts-to-board-
      of-directors-4980",
    "title" : "Canfor appoints Dianne Watts to board of directors",
    "pub_date" : ISODate("2018-06-08T00:00:00.000Z"),
    "text" : "'We are excited to have Dianne join our board and look forward to the experience,
      knowledge and fresh perspective that she will bring,' said Don Kayne, (...) company."
}
```

Code Snippet 1.2. Sample document of "publications" collection.

Intelligence. At this stage, we applied the MapReduce algorithm to identify the key technologies present in the documents collected. The algorithm is written in JavaScript and considers the ontologies terms and their relationships. A Python function was used to convert the OWL file content into a JSON-based dictionary within the JavaScript code.

The Map function converts the publication texts to lowercase and use a regular expression to look for technology mentions in them. As an output, this function returns the publication date, the founded technology and its superclass, and the number "1", which indicates that at least one term was found in the analyzed document. Then, the Reduce function adds the same terms found and returns the total value associated with the label "term_count".

The execution of this algorithm was done in the MongoDB environment, and generated a new collection of documents called "TechnologyMapping". Each document of this collection has a format as shown in the code snippet 1.3. In it,"pub_date" is the month in which the document was published, "technology" means the key technology found, "superclass" means the ontology class to which the technology belongs, and "value" is the document number in which the technology was mentioned that month.

```
{
    {
        "_id" : {
        "pub_date" : "2017-03-01",
        "technology" : "Composite Wood",
        "superclass" : "Biotechnology"
        },
    "value" : 1.0
    }
}
```

Code Snippet 1.3. Output from the MapReduce Processing.

With the publication date and the amount of technology mentions obtained, we built a monthly time series by highlighting the number of documents in which a key technology appeared over the months. It is known that a process is stationary when the characteristics do not undergo changes over time, i.e. a process that develops randomly in time and oscillates around the mean. The stationary time series is the inverse of a non-stationary time series, in which growth or fall trends occur and, then, the mean and variance are related to time.

In order to know the unit root thesis for each time series constructed (formed by the number of documents that mention a technology monthly), the Augmented Dickey-Fuller Test was performed, which helps in detection of non-stationary time series, and verifies if the series used follow a steady stochastic process. This type of process would indicate that there would be no significant variations in the number of key technologies mentioned, and there could be a mature technology. On the other hand, a non-stationary series may indicate a tendency for growth or a drop in interest in a particular key technology.

Finally, graphs of the series were constructed and a polynomial of order 2 was adjusted, in order to facilitate the visualization and analysis of the specialists. The difference between the values *y0* and *y1* was calculated, representing the number of mentions, in percentage, in order to give an idea of the size of the variation and whether it was positive or negative.

Communication. In order to communicate the results to specialists, the system creates a report containing a summary of the main data, such as the number of documents captured, the ten most mentioned technologies, and the associated charts.

4.2 Experimental Results

During the initial stage of Planning, an OWL file containing the ontologies modeled for the Furniture and Wood sector was produced. This file comprises key technologies organized in the form of sheets. These technologies are (in alphabetic order): 3D printing, Additive Fabrication, Additive Manufacturing, Augmented Reality, Automation, Automation and Robotics, Biomass, Biotech, Biotechnology, Certifications, Composite Wood, Construction Wood, Distinctive Design, Engineered Wood, Hardboard, Health and Safety, High Density Fiberboard, Information and Communication Technology, Information Technology, Liquid Wood, Medium Density Fiberboard, Medium Density Particleboard, Multifunctional Furniture, Nanotechnology, Optimized Furniture, Rapid Prototyping, Robotics, Sensory Design, Single Households, Smart Furniture, Strategic Design, Virtual Reality, Waste Management, and Wood Frame.

In the Collection stage, web crawlers were constructed to scrape publications dated between June 1st, 2017 and June 1st, 2018. Table 3 shows the quantity of publications collected from the portals For Mobile, Wood Business, Furniture News, and Woodworking Network. After collected, these publications were organized and saved as MongoDB documents, as detailed above.

Table 3. Number of publications collected in each portal.

Portal	Publications collected
For Mobile	30
Wood Business	190
Furniture News	198
Woodworking Network	2500
Total	**2.918**

During the Intelligence stage, we synthesized indicators such as the number of documents in which the monitored technologies were mentioned (Table 4) and the volume of documents that quoted each class of the key technologies of the ontology (Table 5). In this way, it is possible to see which key technologies and classes were most cited in the period.

Table 4. Number of documents in which the technologies were cited in the period.

Key technology	Number of documents
Automation and Robotics	132
Wood for building	72
Biomass	31
Composite wood	31
Information and Communication Technologies	28
Virtual Reality	18
Health Safety	12
MDF	8
3D Printing	7
HDF	6
Nanotechnology	6
Augmented Reality	6
Waste Management	3
Strategic Design	1

Further details are presented in the Figs. 4 and 5, which show the total of documents that mentioned each key technology and the ontology-defined classes over the monitored period of time.

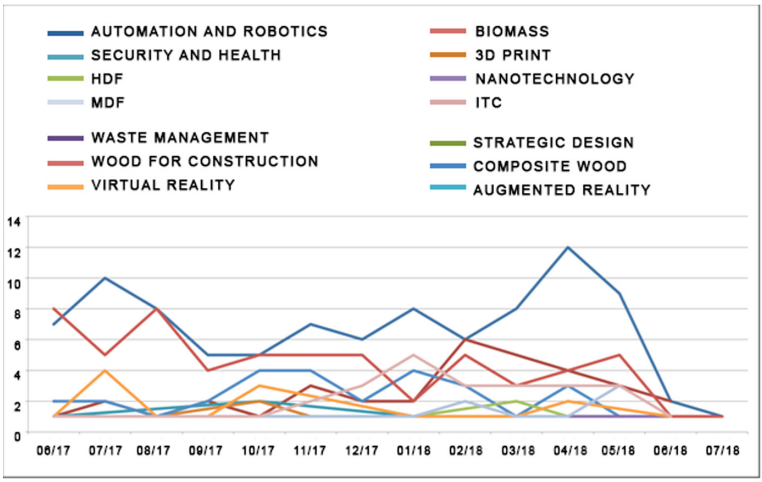

Fig. 4. Total of documents that mention the key technologies, per month.

When analyzing Fig. 4, one can observe that "Automation and Robotics" had a significant increase of citations between February 2018 and April 2018, experiencing a dropping ever since. In both Fig. 4 and 5, it is possible to notice that the volume of citations of most of the terms that oscillate do not show a visible tendency.

There are obvious difficulties related to the evaluation of trends by means of using exclusively the visual graphic presented in Fig. 4. To overcome that, the behavior of each monitored key technology was separately evaluated, in order to verify if its time series presented only oscillations around the same axis, tending to be stationary, or if it would be possible that it might have some kind of tendency associated.

As a tool for supporting specialists, the Augmented Dickey-Fuller Test, was applied. When considering a p-value > 0.05 and accepting the null hypothesis (H0), we concluded that the series has a unit root and is non-stationary. On the other hand, for a p-value $<= 0.05$ we reject the null hypothesis (H0), i.e. indicating that the series has no unit root and is stationary. Thus, it is possible to have a support to assess whether the key technology is becoming popular or not. In Fig. 6, it is possible to see an example of monitoring of Biomass term. The x-axis represents the months. The y-axis represents the count of documents in which the term was identified (Table 6).

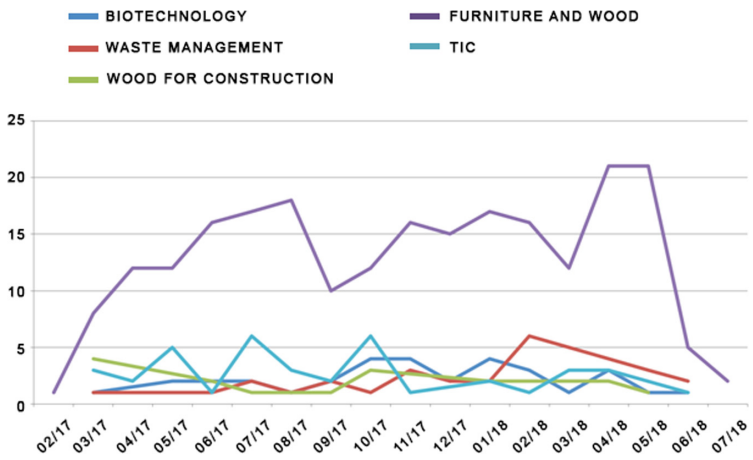

Fig. 5. Total of documents that mention the ontology-defined classes, per month.

Table 5. Number of documents in which the ontology-defined classes are present.

Classes	Number of documents
Furniture and wood	254
Biotechnology	31
Waste Management	31
Information and Communication Technologies	31
Wood for construction	14

Table 6. Augmented Dickey-Fuller Test result for the monitored term "Biomass".

General output	Critical values
Variation in Y axis: +65.22%	1%: −4.223
ADF Test: −2.453971	5%: −3.
p-value found: 0.127079	10%: −2.730
p-value > 0.05: It accepts the null hypothesis (H0); the data has a unit root and is non-stationary	

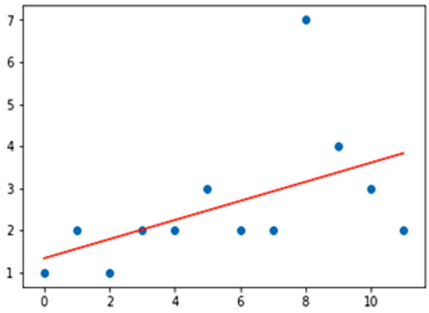

Fig. 6. Number of documents with 'Biomass' term (x: months; y: total of docs).

5 Final Considerations

This paper proposed an automated method for the identification of key technologies in specialized information sources. The proposed method is based on the Technological Surveillance methodology of Sánchez Torres and Palop Marro (2002) [22], and comprises 5 steps: Planning, Collecting, Organization, Intelligence, and Communication. A prototype was developed in order to validate this proposal. Key technologies of the Furniture and Wood sector were identified and quantified from 2,918 publications collected from four specialized Web portals, suggested by Furniture and Wood specialists.

The proposed method for identifying the use of existing technologies in Big Data scenarios, by means of the structured reports, graphs and tables generated by it, is the answer to our research question. That is, now we can automatically identify a set of technologies of interest with increasing popularization in Web portals.

As for future works, we suggest the creation of a sub-step of the Collection phase, which is capable of recommending new possible key technologies to be monitored based on text mining techniques. Although this feature would make the process richer, it would not rule out the human intervention. Another possible evolution would be to extract geographic data from the collected data, in order to check the dispersion of technologies on the map, as well as their intersection with data come from other sources of information.

References

1. Abe, H., Tsumoto, S.: Detecting temporal trends of technical phrases by using importance indices and linear regression. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS (LNAI), vol. 5722, pp. 251–260. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04125-9_28

2. ABNT. Brazilian Association of Technical Standards. Guidelines for research, development and innovation management systems (R&D&I) [In Portuguese: Associação Brasileira de Normas Técnicas. Diretrizes para sistemas de gestão da pesquisa, do desenvolvimento e da inovação (PDI).]. https://www.abntcatalogo.com.br/norma.aspx?ID=088796. Accessed 28 Oct 2020
3. AENOR. Spanish Association for Standardization and Certification. Spanish Experimental Standard Rule UNE 166006 R&D&I: Technological Surveillance System [In Spanish: Asociación Española de Normalización y Certificación. Norma Española Experimental UNE 166006 Gestión de la I +D +i: Sistema de Vigilancia Tecnológica]. https://www.aenor.com/normas-y-libros/buscador-de-normas/une?c=N0059973. Accessed 28 Oct 2020
4. Salgado Batista, D., Guzmán Sánchez, M. V., Carrillo Calvet, H.: Establishment of a scientific-technological surveillance system. Technol. Forecasting [Original title in Spanish: Establecimiento de un sistema de vigilancia científico-tecnológica]. ACIMED 11(6) (2003)
5. Buckland, M.K.: Information as thing. JASIS **42**(5), 351–360 (1991)
6. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
7. Ena, O., Mikova, N., Saritas, O., Sokolova, A.: A methodology for technology trend monitoring: the case of semantic technologies. Scientometrics **108**(3), 1013–1041 (2016). https://doi.org/10.1007/s11192-016-2024-0
8. Flick, U.: Introduction to Qualitative Research [Original title in Portuguese: Introdução à Pesquisa Qualitativa]. Artmed, Porto Alegre (2009)
9. Geum, Y., Jeon, J., Seol, H.: Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing. Technol. Anal. Strategic Manag. **25**(1), 1–22 (2013)
10. Gil, A.C.: How to design research projects [Original title in Portuguese: Como elaborar projetos de pesquisa]. Atlas, Barueri (2010)
11. Hakim, A.R., Djatna, T.: Extraction of multi-dimensional research knowledge model from scientific articles for technology monitoring. In: 3rd International Conference on Adaptive and Intelligent Agroindustry (ICAIA), pp. 300–305. IEEE (2015)
12. Hjørland, B., Albrechtsen, H.: Toward a new horizon in information science: domain analysis. J. Am. Soc. Inf. Sci. **46**(6), 400–425 (1995)
13. Zikopoulos, P., Eaton, C.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, New York City (2011)
14. Muñoz Durán, J., Marín Martínez, M., Vallejo Triano, J.: Technological surveillance in R&D&I project management: resources and tools [Original title in Spanish: "La vigilancia tecnológica en la gestión de proyectos de I +D +i: recursos y herramientas"]. El profesional de la información **15**(5), 411–419 (2006)
15. Laurila, J., et al.: The mobile data challenge: Big data for mobile computing research. Nokia Research Center (2012). https://www.academia.edu/20092648/The_mobile_data_challenge_Big_data_for_mobile_computing_research. Accessed 27 Oct 2020
16. León, A.M., Castellanos, O.F., Vargas, F.A.: Assessment, selection and relevance of software tools used in technological surveillance [Original title in Spanish: Valoración, selección y pertinencia de herramientas de software utilizadas en vigilancia tecnológica]. Ingeniería e investigación **26**(1), 92–102 (2006)
17. Momeni, A., Rost, K.: Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. Technol. Forecast. Soc. Change **104**, 16–29 (2016)

18. Nam, S., Kim, K.: Monitoring newly adopted technologies using keyword based analysis of cited patents. IEEE Access **5**, 23086–23091 (2017)
19. Neighbors, J. M.: Software construction using components. University of California, Irvine (1980). https://escholarship.org/content/qt5687j6g6/qt5687j6g6.pdf. Accessed 28 Oct 2020
20. OVTT. Virtual Observatory for Technology Transfer. Technology Surveillance Concept [In Portuguese: Observatório Virtual de Transferência de Tecnologia. Conceito De Vigilância Tecnológica]. https://pt.ovtt.org/vigilancia-tecnologica-conceitos. Accessed 28 Oct 2020
21. Palop, F., Vicente, J.M.: Technological Surveillance and Competitive Intelligence. Its potential for Spanish companies [Original title in Spanish: Vigilancia tecnológica e inteligencia competitiva: su potencial para la empresa española]. Cotec, Madrid (1999)
22. Sánchez Torres, J.M., Palop Marro, F.: Software tools for the practice in the company of Technological Surveillance and Competitive Intelligence [Original title in Spanish: "Herramientas de software para la práctica en la empresa de la Vigilancia Tecnológica e Inteligencia Competitiva"], Evaluación Comparativa, 1ª Edición. TRIZ, España (2002)
23. Park, H., Kim, E., Bae, K.J., Hahn, H., Sung, T.E., Kwon, H.C.: Detection and analysis of trend topics for global scientific literature using feature selection based on gini-index. In: IEEE 23rd International Conference on Tools with Artificial Intelligence, pp. 965–969. IEEE (2011)
24. PDIC2022. Industry Development Program of the Brazilian State of Santa Catarina for 2022: competitiveness with sustainability [In Portuguese: Programa de Desenvolvimento da Indú'stria Catarinense 2022: competitividade com sustentabilidade]. http://www4.fiescnet.com.br/homepdic. Accessed 28 Oct 2020
25. Sharma, S., Gupta, P.: The anatomy of web crawlers. In: International Conference on Computing, Communication Automation, Greater Noida, pp. 849–853. IEEE (2015)
26. Shiryaev, A.P., Dorofeev, A.V., Fedorov, A.R., Gagarina, L.G., Zaycev, V.V.: LDA models for finding trends in technical knowledge domain. In: IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 551–554. IEEE (2017)
27. Smiraglia, R.P.: The Elements of Knowledge Organization. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09357-4
28. Tomaél, M.I.: Information sources on the internet [Original title in Portuguese: Fontes de informação na internet]. EDUEL, Londrina (2008)
29. Wei, Y.M., Kang, J.N., Yu, B.Y., Liao, H., Du, Y.F.: A dynamic forward-citation full path model for technology monitoring: an empirical study from shale gas industry. Appl. Energy **205**, 769–780 (2017)
30. White, T.: Hadoop: The Definitive Guide. O'Reilly, Beijing (2015)