Journal of Intelligence Studies in Business

Journal of Intelligence Studies in Business

Publication details, including instructions for authors and subscription information: https://ojs.hh.se/index.php/JISIB/index

Intelligent information extraction from scholarly document databases

Fernando Vegas Fernandez[a]*

[a]Departamento de Ingeniería Civil: Construcción, Universidad Politécnica de Madrid, Spain; *fvegas@ciccp.es

To cite this article: Vegas Fernandez, F. (2020) Intelligent information extraction from scholarly document databases. *Journal of Intelligence Studies in Business*. 10 (2) 44-61.

Article URL: https://ojs.hh.se/index.php/JISIB/article/view/570

PLEASE SCROLL DOWN FOR ARTICLE

This article is Open Access, in compliance with Strategy 2 of the 2002 Budapest Open Access Initiative, which states:

Scholars need the means to launch a new generation of journals committed to open access, and to help existing journals that elect to make the transition to open access. Because journal articles should be disseminated as widely as possible, these new journals will no longer invoke copyright to restrict access to and use of the material they publish. Instead they will use copyright and other tools to ensure permanent open access to all the articles they publish. Because price is a barrier to access, these new journals will not charge subscription or access fees, and will turn to other methods for covering their expenses. There are many alternative sources of funds for this purpose, including the foundations and governments that fund research, the universities and laboratories that employ researchers, endowments set up by discipline or institution, friends of the cause of open access, profits from the sale of add-ons to the basic texts, funds freed up by the demise or cancellation of journals charging traditional subscription or access fees, or even contributions from the researchers themselves. There is no need to favor one of these solutions over the others for all disciplines or nations, and no need to stop looking for other, creative alternatives.

Journal of Intelligence Studies in Business
Vol. 10, No. 2 (2020) pp. 44-61
Open Access: Freely available at: https://ojs.hh.se/

Intelligent information extraction from scholarly document databases

Fernando Vegas Fernandez[a]*

[a]*Departamento de Ingeniería Civil: Construcción, Universidad Politécnica de Madrid, Spain*
**Corresponding author: fvegas@ciccp.es*

Received 4 January 2020 Accepted 5 May 2020

ABSTRACT Extracting knowledge from big document databases has long been a challenge. Most researchers do a literature review and manage their document databases with tools that just provide a bibliography and when retrieving information (a list of concepts and ideas), there is a severe lack of functionality. Researchers do need to extract specific information from their scholarly document databases depending on their predefined breakdown structure. Those databases usually contain a few hundred documents, information requirements are distinct in each research project, and technique algorithms are not always the answer. As most retrieving and information extraction algorithms require manual training, supervision, and tuning, it could be shorter and more efficient to do it by hand and dedicate time and effort to perform an effective semantic search list definition that is the key to obtain the desired results. A robust relative importance index definition is the final step to obtain a ranked importance concept list that will be helpful both to measure trends and to find a quick path to the most appropriate paper in each case.

KEYWORDS Business intelligence, concept map, information extraction, knowledge management, literature review, natural language process, NLP, semantic search

1. INTRODUCTION

According to the Cambridge dictionary, knowledge is "understanding of or information about a subject that you get by experience or study, either known by one person or by people generally". It could also be defined as "the state of knowing about or being familiar with something" or "the creation of information from structured or unstructured data" (Upadhyay and Fujii 2016). In other words, knowledge is the result of settling information. "The general purpose of knowledge discovery is to extract implicit, previously unknown, and potentially useful information from data" (Matsuo and Ishizuka 2004).

Information can be contained in a lot of documents available in several kinds of formats (Mitra and Chaudhuri 2000), as can be seen in Figure 1. Nowadays there is no distinction between electronic and printed formats given that any printed paper can be easily converted to an electronic format with scanning and OCR technologies that are commonplace.

A large amount of available information on the Internet has made it easier to reach a constantly increasing number of documents but it has caused the problem of finding the most relevant ones for the specific purpose that the user addresses. Information retrieval (IR) has attracted scientists' attention since the 1960s (Allan et al. 2002). Allan uses Salton's definition in 1983 for IR: "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information". Recent publications define IR as "A system to identify a subset of

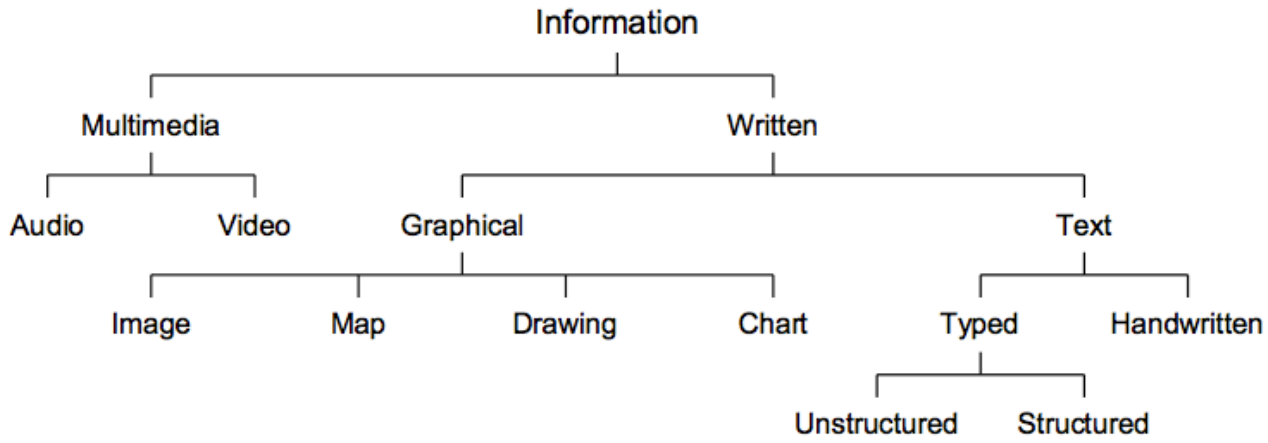

Figure 1 Distinct information formats.

documents in a large text database or a library scenario a subset of resources in a library" (Grishman 2019).

An information extraction system identifies a subset of information within a document to extract relevant information from documents. Information extraction (IE) should not be confused with the more mature technology of information retrieval (IR) (Gaizauskas and Wilks 1998). To sum it up, IR retrieves relevant documents from collections and IE extracts relevant information from documents. The relevance of extracted information is always related to the interests, goals, and specific information requirements of the researcher and, then, once it has been internally processed, information becomes knowledge.

Extracting knowledge from big databases and document databases has long been a challenge because of the large number of documents that make it hard to select the most relevant data. For that reason, a lot of retrieval algorithms have been developed (Ahmad and Ansari 2012; Boden et al. 2012; Karol and Mangat 2013; Koval and Návrat 2012; Wang et al. 2013) applying distinct sophisticated techniques: fuzzy, artificial neural network (ANN), clustering, machine learning, and hybrids.

There is a specific scenario where the challenge is not to find the right documents but to extract usable information from them: it is the literature review that every researcher faces when addresses a new research project (Nasar et al. 2018). This is a case of unstructured typed text written information (see Figure 1). In that situation, IR can be easily solved with the available search engines on the Internet. However, it is much harder to extract and manage information because a very high accuracy is needed and information about many distinct concepts should be extracted from documents depending on the researcher's requirements. In that scenario, knowledge management involves not just information about keywords, tags, and meta-data, but a structured and even quantitative structure of all the concepts that can be relevant for the researcher's objectives.

The document database size that researchers use in each specific research project is very small, typically 30 to a few hundred documents, and this situation is far from big data scenarios. For that reason, most of the time and effort should be dedicated to clearly defining specific user information requirements before thinking of a better way to extract information.

This article addresses the case of the literature review. Researchers do a literature review, create a document database, and must manage that source of knowledge. There are several tools to manage that kind of document (e.g., EndNote, Mendeley, Word), but they just provide a catalog management functionality, When it comes to extracting knowledge, there is a severe lack of functionality. This case is a "little brother" of the general problem of extracting information from PDF files, but the approach, methodology, and principles used in this case are the same as those used in bigger cases. However, the IT tools required are much simpler.

Before searching for concepts in a document database (e.g., ideas, topics) it is necessary to perform a previous concept analysis to define the semantic framework that will be used later (López-Robles et al. 2019; Sarwar and Allan 2019). Sometimes this analysis can be easily performed because it merely consists of defining words to be found in the text (e.g., to achieve a list of possible risks) and other times

it is harder. This article proposes a simple and effective way to extracting information from research document databases depending on the researcher's predefined breakdown structure, obtaining a ranked list of concepts and items to define priorities or to make decisions. These results are relevant for researchers and are an example of what companies could do to organize and use their stored information simply and effectively.

2. PROBLEM DESCRIPTION

Researchers use literature review as a relevant part of their research studies to know the state of the art and to give a sound basis to the statements they include in their papers. Each new research project leads to a new tailored document database creation with a few hundred documents that, although possibly partially overlapping with previously used databases, is a fully new one from which researchers will take references to include them in their new papers. In fact, they create a library that could be seen as their business intelligence document warehouse (Tseng and Chou 2006), because researchers do not use their document database just to cite previous works but also to extract knowledge from those documents.

Scholarly documents address a specific subject and give a conclusion. Researches can read abstracts and even write a summary for each document. But there is much more information there, related to the main subject and related to marginal topics that might concern researchers, for which they might need to keep a record by annotating statements, methods, algorithms, author's position about specific issues and techniques (Rostami et al. 2015). To do that, researchers could think of a predefined information breakdown structure and a list of premises, concepts, ideas, issues, and techniques that they would like to confirm or refute with the database information. In the end, that's knowledge (Sirsat et al. 2014), and that sort of virtual list containing a reduced number of entries (typically 20 to 50) is itself a handy knowledge reference.

Researchers need tools to efficiently carry out that task, but they usually do it by hand or with the help of desktop cataloging tools such as EndNote, Mendeley, or Word. A survey conducted in Universidad Politécnica de Madrid with a selected group of Ph.D. candidates and researchers confirmed this statement. Sophisticated algorithms are not always the right answer to extract information and knowledge, and most researchers are not opened to them because they do not have enough time to try them. Furthermore, most of the scholarly algorithms proposed require manual training, supervision, and tuning (Sirsat et al. 2014; Upadhyay and Fujii 2016) and, in the end, it is faster and more efficient to do it by hand.

Researchers need to retrieve information from scholarly papers and transform it into knowledge. A possible way is to create a list of concepts or items that are representative of each document concerning what researchers are looking for in their research projects. That list of concepts can be weighted later on to achieve a ranked list of relevant concept elements with the overall reviewed literature.

3. OBJECTIVE

This article addresses the literature review and the knowledge extraction that researchers carry out using scholarly document databases in their research projects and aims to give an affordable solution to improve that situation. Scientific document databases are much more than a collection of papers that need to be managed and cataloged: a task that several commercial solutions can do. Scientific document databases are a relevant source of information and researchers need to extract knowledge from them and rank results according to their relevance.

4. RESEARCH METHOD

This study analyzes the state of the art in intelligence information extraction from scientific document databases. To do that, a systematic literature review and interviews with researchers at Universidad Politécnica de Madrid were carried out. That way, requirements and available resources were identified. This study also takes advantage of my personal experience as a researcher and as a Chief Information Officer in multinational companies.

Advances in linguistic structure definitions were studied in depth to try to find the most efficient way to analyze text and to use it for specified purposes. Novelty proposed algorithms were considered to evaluate their adequacy for the objectives proposed.

A previous author's experience related to a competitive intelligence innovation project studied in 2015-2016 to predict risks in projects is a significant reference as to what actual technical solutions can provide and their

possibilities to satisfy the requirements proposed in this study.

5. LITERATURE REVIEW

A systematic literature review was performed to know the state of the art related to intelligent information extraction following the searching method by Bettany-Saltikov (Bettany-Saltikov 2012; Kasperiuniene and Zydziunaite 2019; Snyder 2019). A systematic search, unlike a narrative search that could yield a subset of haphazard and biased documents, achieves a neutral collection of documents to obtain an objective view of the state of the art.

To carry out the information retrieval, the initial idea of using the string "intelligent information extraction" linked to scholarly and scientific documents was completely dismissed because it hardly gave any results; a search for the concept "intelligent information extraction from document databases" was performed in several sources (Renault and Agumba 2016; Xia et al. 2018), with and without quotation marks and sometimes splitting that string into smaller fragments to achieve complementary results. As some sources retrieved more than 313,000 documents (e.g., Google Scholar), the first 400 hits were selected in each source, given that their search engines are supposed to show the most relevant results first. That outcome was filtered screening titles, keywords, and abstracts to rule out documents that did not meet the subject proposed and those that were unreachable.

The results obtained prove that distinct sources do not always contain distinct databases; their search engines are different, and, for that reason, their first documents retrieved were distinct. It is possible to find in Google Scholar almost any document found in the other sources. However, by using distinct sources it is possible to get more results. The number of remaining documents, after filtering and deleting duplicated results, was 58.

Concepts such as natural language processing, semantics, and ontologies frequently appear in the documents reviewed. A linguistic approach to the ontology concept could be helpful to clarify its meaning with several distinct definitions (Schalley 2019): "An explicit specification of a conceptualization", "The study of the categories of things that exist or may exist in some domain", and "Catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D".

Some documents address only IR (Allan et al. 2002; Barde and Bainwad 2018), others only address IE (Lee 1998; Saik et al. 2017), and most of them address both IE and IR. Although IE and IR have been studied from the 1960s, there is a lack of scholarly documents addressing IE and IR from scientific publications: only 7 out of the 58 documents retrieved address them (Esposito et al. 2005; Marinai 2009; Nasar et al. 2018; Rodríguez et al. 2009; Saik et al. 2017; Upadhyay and Fujii 2016; Wang et al. 2013):

Esposito addresses a semantic-based tag extraction by using their system DOMINUS, and they achieve accuracies from 93% up to 98% (Esposito et al. 2005). However, those tags are title, author, abstract, and references, and nowadays it is easier to retrieve those tags with Google Scholar and tools such as EndNote and Mendeley.

Marinai aims to extract administrative meta-data from digital articles (Marinai 2009). The paper uses the term "administrative meta-data" to describe details such as title, authors, and publisher (named hereinafter "administrative tags" to avoid confusion). Their outcome is, thus, a file card, the sort of data that tools such as EndNote and Mendeley can provide.

Nasar et al.'s article distinguishes meta-data extraction and key-insights extraction and says that "the amount of time that is required to conduct a quality review can take up to 1 year" and that a "systematic literature review can take up to 186 weeks with single/multiple human resources". In the survey, they talk about an average accuracy of 92% in retrieving meta-data when the document includes a Report Document Page and 64% when it does not. When it comes to key-insight extraction, the precision is 42% and the recall is 52% (Nasar et al. 2018).

Rodríguez et al. wrote in 2009 a promising article trying to classify software engineering publications with a three-step method using natural language processing (NLP), mainly focused on (but not limited to) HTML documents. No information is provided about their results, precision, and recall rates (Rodríguez et al. 2009).

Saik et al.'s article addresses the agricultural biotechnology field to automatically extract medical and biological knowledge from the PubMed texts using semantic analysis and the relational database

MySQL. They propose the use of an adapted version of their ANDSystem solution that "involved the creation of a subject domain ontology and semantic linguistic rules (templates) for analyzing natural language texts and extracting knowledge formalized according to a given ontology". It requires "dictionaries of the objects" that must be first created using templates (Saik et al. 2017).

Upadhyay and Fujii propose "a practical sentence extraction procedure and supporting system which we intended to call knowledge extraction system" by applying rules to identify and extract keywords, discourse keywords, and sentences, but human expert support is required and no precision nor recall rates are provided (Upadhyay and Fujii 2016).

Wang et al. focus on information retrieval (document retrieval) based on word concepts and text clustering. They apply the COSINE algorithm to classify documents (Wang et al. 2013).

Natural language processing (NLP) is a constant reference in most publications (Hassan and Le 2020). Sometimes their proposals ask for structured documents and, when not, they need to transform documents into structured data (Dezsenyi et al. 2007; Oro and Ruffolo 2008). Other times they need to convert the original PDF files into HTML and text format files to be able to proceed (Hassan and Baumgartner 2005a; Rizvi et al. 2018; Seng and Lai 2010). The methods and algorithms proposed frequently require the involvement of experts and manual training and tuning of the system (Chen and Lynch 1992; Koval and Návrat 2012; Lambrix and Shahmehri 2000; Sirsat et al. 2014; Upadhyay and Fujii 2016).

The documents analyzed propose algorithm-based systems and agents with rules to query document databases, although it is common to find unsolved problems when there are heterogeneous data sources (Seng and Lai 2010). Sometimes the solution proposed is just a query with Boolean logic (Lambrix and Shahmehri 2000; Lee 1998; Rahman et al. 2017; Sarwar and Allan 2019) and other times they propose sophisticated techniques such as an artificial neural network (Al-Hroob et al. 2018; Matos et al. 2010), machine learning (Fan et al. 2015; Hassan and Le 2020; Seedah and Leite 2015), and artificial intelligence (Ansari et al. 2016; Gupta and Gupta 2012; Matsuo and Ishizuka 2004), even though artificial intelligence is usually related to NLP (Kim and Chi 2019; Lee 1998).

Some documents address information extraction from multimedia contents and files (Srihari et al. 2000; Wolf and Jolion 2004). Other works are intended for specific purposes such as biological knowledge extraction from biomedical web documents (Hu et al. 2004), medical document summarization (Afantenos et al. 2005), and software testing (Lutsky 2000). Some studies aim for "automatic keyword extraction" by considering co-occurrence and frequency to extract keywords (Matsuo and Ishizuka 2004), but do not consider the researcher's interests.

Clustering and classifying techniques are often used, such as nearest neighbor classifier, Bayes, and support vector machine (Shrihari and Desai 2015; Song et al. 2007). Attempts to intelligently split unstructured PDF files into segments have been made by using ontologies and queries to generate an XML output with understandable data, trying to simulate how human readers would analyze a page (Hassan and Baumgartner 2005b). That "human visual" approach has also been addressed by other authors trying to make text visual, although there is a generalized lack of references and there are strong limitations (Nualart-Vilaplana et al. 2014).

There are many proposals although sometimes they have not been fully tested (Inui et al. 2008) and are just experimental proposals (Fan et al. 2015; Karthik et al. 2008; Li et al. 2015; Milward and Thomas 2000; Xie et al. 2019). The most frequent situation is that the systems proposed need human training, supervision, and tuning (Fan et al. 2015; Sirsat et al. 2014; Upadhyay and Fujii 2016), and even with that, the outcome is not always as good as desired, with poor precision and recall values (Adrian et al. 2015; Al-Hroob et al. 2018; Milward and Thomas 2000).

6. PROPOSED APPROACH

In this section, several relevant components of the whole problem are analyzed, creating a breakdown structure to address them separately.

The typical path that researchers follow in their literature review process has several stages (Xia et al. 2018). According to Xia, there are three stages: stage 1 includes review planning and searching for relevant articles using electronic databases; stage 2 involves deleting all duplicates according to the title and author and excluding irrelevant papers by reading their titles, abstracts, and keywords; and stage 3 refers to content analysis. We

propose a more effective procedure with four stages (Figure 2).

6.1 Stage 1: planning and computer search

In stage 1 an electronic search is performed using databases and search engines on the Internet. To do that, a previous selection of databases is done considering the research subject, e.g., Google Scholar, Web of Sciences, Scopus, or ResearchGate. Some of those databases share documents: that means that they could have the same content, although the result of the search performed can be quite different because of their different search engines. It is relevant to notice that Google Scholar contains almost every reference included in the other databases, and Stage 3 will take advantage of this fact to automatically obtain document tags.

After having selected the desired databases, it is necessary to define the keywords and patterns that will be used with the search engines selected. As it is very easy to perform search operations, it is possible to use several keywords and patterns, with and without quotation marks and sometimes splitting search strings into smaller fragments to achieve complementary results.

With each search operation, the outcome is a list of documents that match the query. When the number of results is too high it is necessary to refine the search by changing the keywords and patterns or to select just the desired number of results. Those outcomes can be easily copied and pasted into a spreadsheet,

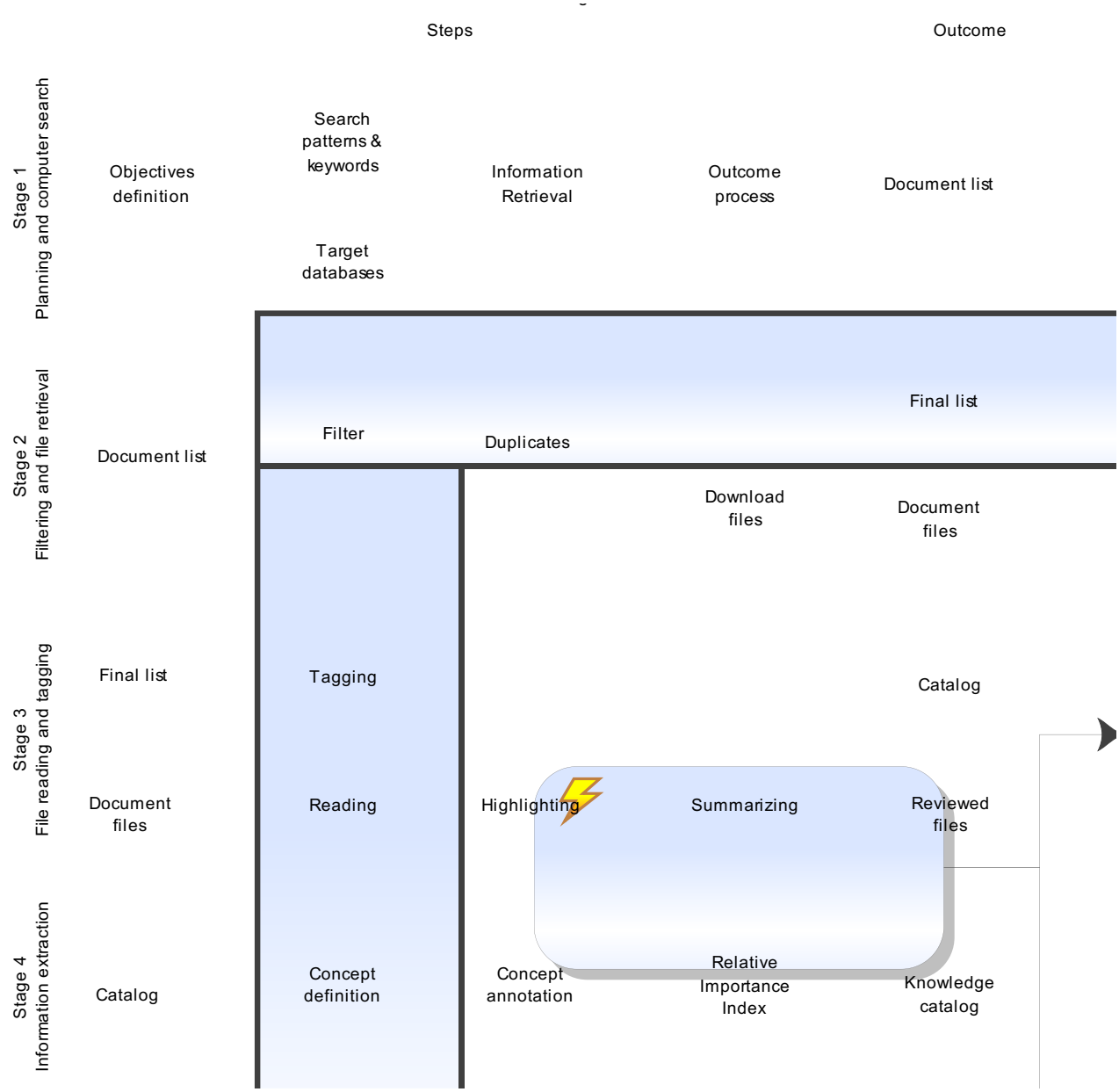

Figure 2 Process stage description.

like Excel, to transform them into easy to use reports. Depending on each database, those lists could contain a variable number of identification fields such as title, authors, date, and even abstract and other tags ("administrative tags"). All that information can be used in stage 2 for filtering purposes.

The feasibility, agility, and flexibility of modern search engines lead to dismissing, in general, any other possible sophisticated algorithm proposed in the IR literature.

6.2 Stage 2: filtering and file retrieval

In stage 2 a filtering operation is performed to refine the results obtained in the previous stage. Excel filters are used to select or unselect document titles to exclude irrelevant documents. For instance, a possible exclusion rule could be to find in the title the words "image", "video", and "media". Additional available information, e.g., keywords, abstract, or other data, can be used to exclude, for instance, documents corresponding to patents: in this case, the filtering rule would be to find the word "patent" close to the title line. If necessary, documents can be downloaded to check their content and decide whether they fit the subject proposed.

When the filtering operation is completed, duplicate results are detected according to the title and authors and then deleted. Finally, the documents are downloaded, and all unreachable documents are excluded. The outcome of this stage is a final list of documents and a database with downloaded PDF files.

6.3 Stage 3: file reading and tagging

In stage 3, documents retrieved should be tagged and reviewed. Meta-data in scientific documents is information commonly associated with administrative properties, such as author names, title, publication date, or journal (Esposito et al. 2005; Marinai 2009; Tseng and Chou 2006), and many researchers have tried to find ways to retrieve them automatically, even recently (Nasar et al. 2018). However, tagging files is very easy now because it can be done using free tools. For this reason, other possible equivalently sophisticated algorithms proposed in the IR literature were dismissed for this purpose. The most direct way to do it is to look for the document title on Google Scholar and to export the reference obtained to Mendeley, EndNote, or another catalog tool (not all of them are free). Both Mendeley and EndNote are desktop tools to catalog references and to allow researchers to include citation and a reference list properly formatted in their papers. With those tools it is also possible to edit tags and update them automatically. Tags considered in this step are only administrative properties, not other content-related tags (López-Robles et al. 2019; Xie et al. 2019).

All documents are read at this stage and researchers begin to achieve knowledge. According to Xia, "the technique of content analysis is employed for compressing many words of text in an organized manner, identifying the focus of subject matter, and diagnosing emerging patterns in the current body of knowledge" (Xia et al. 2018). The researchers interviewed in Universidad Politécnica de Madrid had distinct ways and tools to carry out paper revision, but highlighting and summary elaboration are a constant for all of them.

At this stage, the action proposed is a revision of the papers with highlighting of parts of the text using different colors and even writing a short summary (about 150 words) with keywords, tips, and short sentences. This summary is not an abstract summary, but a cue to help them to recall document content later on.

6.4 Stage 4: knowledge extraction

According to Hobbs, "Information extraction is the process of scanning text for information relevant to some interest" and "it requires deeper analysis than key word searches" (Hobbs 2002). Natural language process goes beyond the exact term-matching technique (Rahman et al. 2017) and focuses on concepts, semantics, and relationships between terms to try to retrieve most of the original ideas expressed by document writers. It is a hard task for algorithms and programmers to handle entities, relationships, and events to process them automatically with a high level of both precision and recall, and they frequently require human-supervised help (Grishman 2019). However, that task is the daily work of the human brain: every time a person reads a paper, they unconsciously create a mind map which connects the most relevant concepts with their interests to generate knowledge. That virtual mind map could be explicitly created by defining key concepts corresponding to the concepts identified after having analyzed the relevant syntagmas, ontologies, and keywords existing in the text studied (Buzan 2004).

The criteria to define those key concepts is not the frequency-based traditional model (Fan et al. 2015; Matsuo and Ishizuka 2004), but a

tailored definition that researchers can make according to three factors (Sirsat et al. 2014): 1) the overall contribution of the documents studied to the research project, with concepts that attract researcher's attention because they appear in several documents of the database studied; 2) the researcher's previous knowledge that makes them search for specific concepts to clarify authors' position about them; and 3) the researcher's experience, which helps them find concepts that could become relevant according to their perception. Some authors call them "keywords" and "discourse words" (Upadhyay and Fujii 2016). This step affects the final outcome and is directly related to the research project purposes (see Figure 3).

The aim of defining those concepts is not to summarize documents but to summarize their contribution to the research project, making it possible to characterize documents as a sort of layout and schematic summary in the same line followed by some proposals for document image layout analysis (Oliveira and Viana 2017).

According to this, several distinct possible concept types are shown in Table 1. In this table, "type" refers to the way the concept is found in the text reviewed and how it is annotated. Regarding the way to find them ("trigger"), there are two main possibilities: to be a word (or group of words) or to be a sentence. It is a word (or group of words) when their occurrence undoubtedly means a concept expression, e.g., "ANN", and it is a sentence when concepts are expressed in a more complex way so that no single word is enough to summarize those concepts. Regarding the way concepts may appear ("variation") they could be specific words and groups of words or an opened or closed name list. Regarding the way concepts are "annotated" in each document, they can be registered just with an "x" mark (they meet the required keyword, idea, or condition) or they can be labeled with a descriptive list element or name. Last, concepts can be numeric values; in that case, the value is annotated. To fully understand Table 1 a detailed description of the types is included in Table 2.

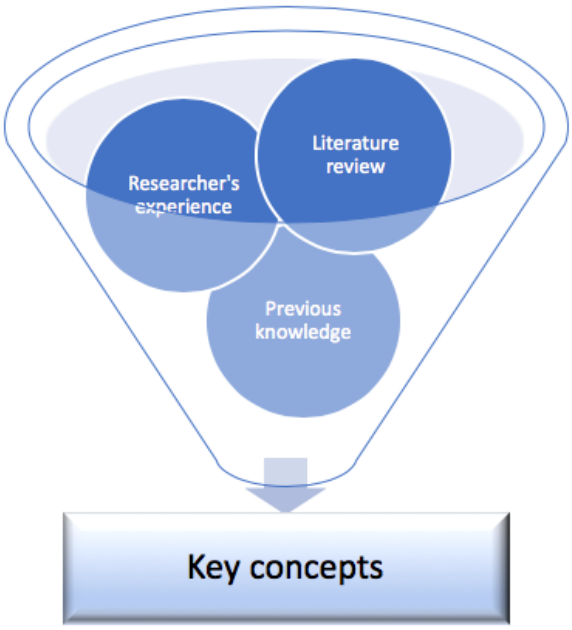

Figure 3 Key concept definition.

Researchers can define as many concepts as needed to cover each detail that is relevant for their research and that they will want to include in their papers. Semantic analysis is an undeniable requirement to achieve a good annotation that is the basis of a key concept definition (Malik et al. 2010).

Once the concept definition has been done, a new document review would be needed to identify them in all the documents and to annotate their occurrences. This operation becomes shorter than it could be thought by using desktop tools that make the use of complicated algorithms and programs unnecessary. There are free solutions, such as Adobe Reader and DocFetcher. DocFetcher creates and uses an internal index (the same

Table 1 Concept types.

Type	Trigger	Variations	Annotation
Keyword	Word	Word, group of words	"x"
Idea/opinion/statement	Sentence	N.A.	"x"
Position	Sentence	N.A.	List element
Use case	Sentence / table / figure	List	List element
Name	Sentence	List	Name
Numeric	Sentence / table / figure	N.A.	Value
Condition	Sentence	List	"x"

Table 2 Type definition.

Type	Definition
Keyword	Applies to the undeniable meaning of a word and group of words in a specific context, e.g., Information Retrieval, Cosine, Query, Machine Learning, Ontology, ANN, or NLP.
Idea/opinion/statement	Applies to a conceptual meaning that could be expressed with distinct words and sentences, e.g., "Need for improvement", "Knowledge extraction", "lack of objectivity", or "biases".
Position	Applies to statements, case of use, and others where authors show whether they approve, reject, or just cite a particular subject, e.g., in regards to a specific technique, they "use or recommend", they "criticize", or they "cite".
Use case	Applies to distinct options researchers might want to keep track of, such as kind of technology, type of chart, or type of scale.
Name	Applies to concepts that can be registered with their names, e.g., system, country, or activity.
Numeric	Applies to concepts that can be quantitatively measured so that it is possible to register their value, e.g., precision or recall.
Condition	Applies to specific conditions that document scope could accomplish to meet the researcher's interests, e.g., specific industry or country, or specific field.

way as Adobe Acrobat does) that allow users to perform quick Boolean searches for any word and string in a document databases. For instance, to find whether documents indicate that further improvement is needed (an idea/opinion/statement type concept), it would be possible to look for "improve" and "limitation" and retrieve the texts "improving the performance of NLP-based tools" and "there are also practical limitations in rule generation …" (Kim and Chi 2019). However, the text "their sometimes low recall may be compensated by adjusting" (Adrian et al. 2015) and "is prone to several limitations that, in turn, offer opportunities for future research" (Li et al. 2015) would not be retrieved.

This manual process is similar to Li et al.'s, which consists of an automated method to retrieve meta-data (Li et al. 2015). Their process lexicon extraction and task identification method for process mining requires manual task annotation to train a statistical model and yields over 75 % classification accuracy, 70 % precision, and 95% recall.

The method proposed here improves accuracy, precision, and recall up to 100%, and it is not more manually time-consuming than most of the automated methods proposed in the literature.

To efficiently register those knowledge tags, the use of a spreadsheet is suggested. This practice allows for an additional feature: a quantitative measure of the relevance of each concept, i.e., the use of a relative importance index (RII). This idea can be found in many works (Alashwal and Al-Sabahi 2018; Jarkas and Haupt 2015; Nagalla et al. 2018) and for this research project, the solution proposed by Vegas-Fernández was used (Vegas-Fernández 2019; Vegas-Fernández and Rodríguez López 2019).

This method applies a weight to each document that considers the document type (standard or regulation, doctoral thesis, book, indexed journal, lecture source, unindexed journal, master thesis, a website run by a renowned organization, or a standard website). The date and their scope are also considered by adding +0.5 to documents after 2010 and by subtracting 0.5 when they are intended for a specific activity or a particular country. The final score is the weight assigned to each document, which is considered when the document matches a concept (regardless if the annotation is an "x", a name, or a value). The RII is the ratio between the weighted count of documents matching a concept and the maximum value that that weighted count takes for a concept.

The outcome at this stage is a ranked list of key concepts, which is a quantitative outcome of knowledge extraction.

7. KNOWLEDGE EXTRACTION EXAMPLE USING THE PROPOSED SYSTEM

The process of knowledge extraction carried out for this study is explained next to make it easy to understand the scope, possibilities, and limits of the proposed system. Each one of the distinct steps at each stage is described here with data that will allow readers to make their guess about this system.

7.1 Stage 1: planning and computer search

Each researcher is used to searching in scholarly databases, and they choose them according to their preferences. Their previous experience and their knowledge of previous publications related to their research project subject give them the required orientation to select the search strings and the best databases. Searching documents in Google Scholar is a must, but the number of possible retrieved documents can be too high. In this case, the chosen search string was "intelligent information extraction from document databases" without quotation marks to be able to achieve results. That search yielded 313,000 results in Google Scholar, but that outcome was truncated to select just the first 400 most relevant titles.

That systematic search process was conducted in eight sources and 974 documents were originally retrieved from Google Scholar, Web of Sciences, Scopus, ScienceDirect, ResearchGate, ASCE, Elsevier, and Mendeley. Outcomes were post-processed in an Excel workbook to manage each database report; that process consisted of converting the HTML information yielded by each search engine into understandable and easy to use Excel rows. This step took less than 3 hours. The number of documents retrieved is displayed in Table 3.

Table 3 Information retrieval initial summary (number of documents).

Source	Initial Outcome
Google Scholar	383
Web of Sciences	2
Scopus	85
ScienceDirect	26
ResearchGate	350
ASCE	20
Elsevier	3
Mendeley	105
Total	974

7.2 Stage 2: filtering and file retrieval

This stage involves a heavy task because often it is not possible to know whether a document will be useful without reading it. According to their titles, keywords, and abstracts, it is possible to perform an initial filter to reject those that do not meet the requirements. Some search engines do not provide abstracts and keywords in their outcomes and the filter can only consider titles. In those cases, a first filter was applied removing unwanted documents according to their titles, and the remaining were downloaded to check by skim-reading whether they met expectations.

Each downloaded document finally accepted was saved in the computer library labeling it with the author-title format. This step took about 60 hours and the number of documents finally selected was 58, after adding manually three more documents. Table 4 shows the number of remaining documents after removing duplicates.

There were three types of documents in the list: 62% were journal articles, 36% conference proceedings, and 2% books. Journal article impact distribution is shown in Figure 4.

Table 4 Information retrieval final summary (number of documents).

Source	Initial outcome	Resulting outcome
Google Scholar	383	24
Web of Sciences	2	2
Scopus	85	6
ScienceDirect	26	0
ResearchGate	350	8
ASCE	20	4
Elsevier	3	0
Mendeley	105	11
Others	-	3
Summary	974	58

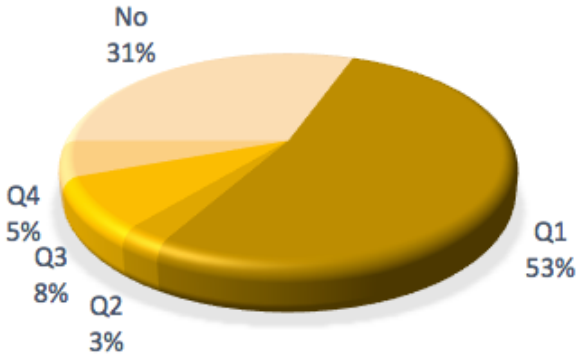

Figure 4 Impact distribution of the retrieved journal articles (Q factor).

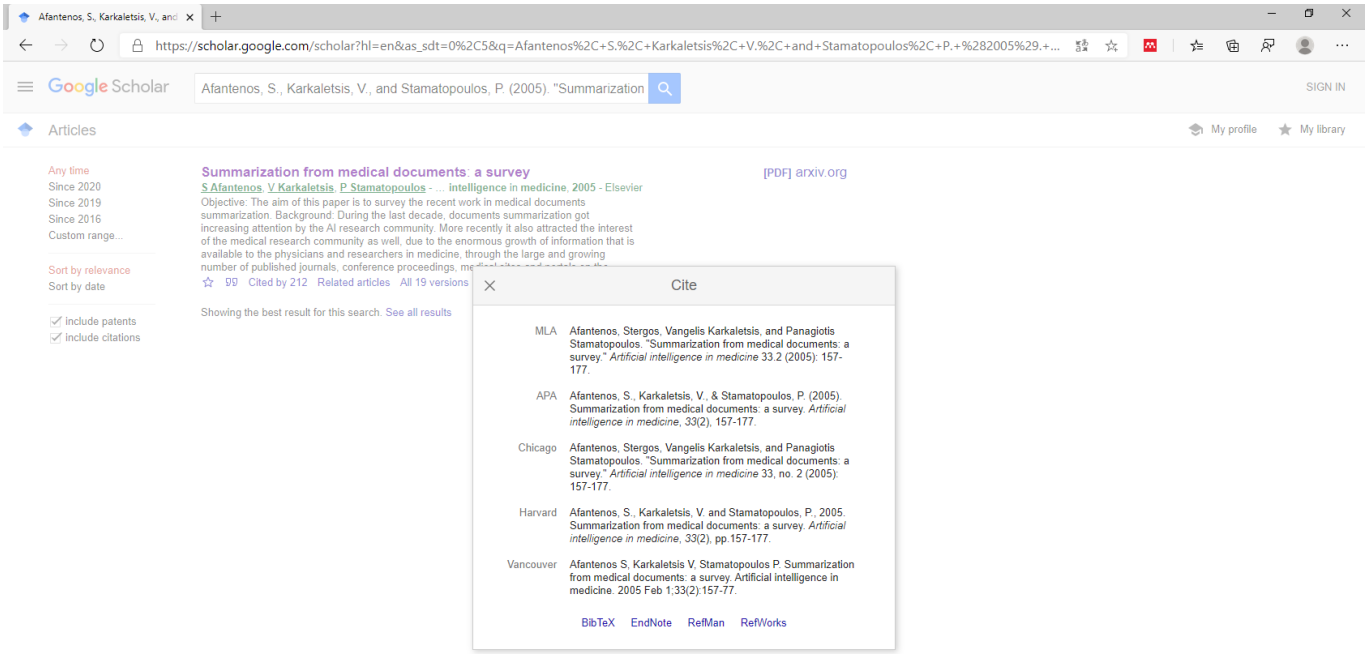

Figure 5 Tag retrieving with Google Scholar.

7.3 Stage 3: file reading and tagging

Two relevant tasks were done at this stage: reading and tagging documents. Google Scholar and its citing tool were used to find each document and to create an entry in the Mendeley catalog (Figure 5).

Most tags are automatically saved, and Mendeley, EndNote, and other tools can find reference updates, although sometimes it is necessary to look for a specific missing tag, such as the DOI, Publisher, or the URL for the document (see Figure 6).

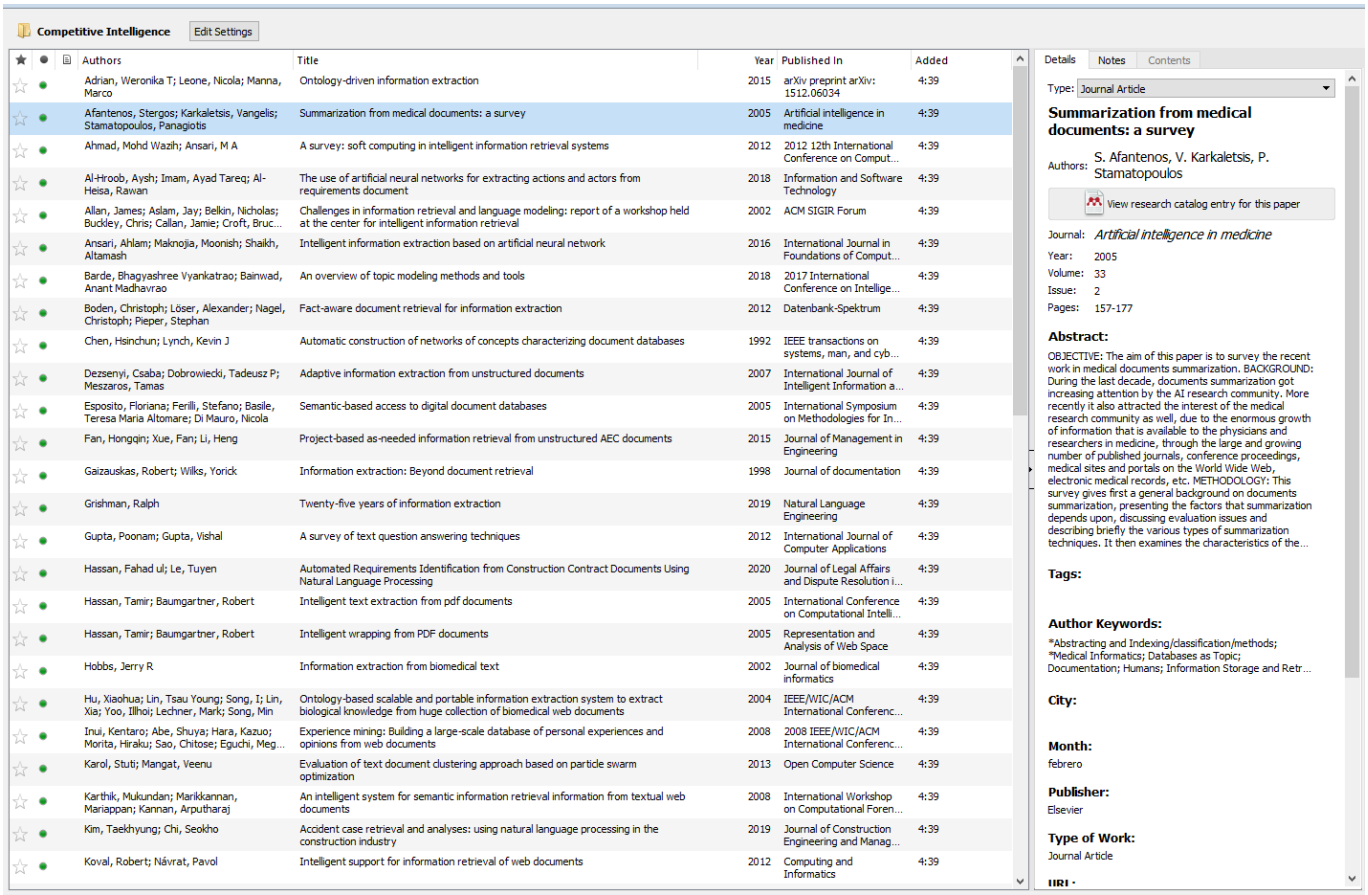

Figure 6 Tag management with Mendeley.

This process does not take long (5 hours for 58 documents), and researchers can perform this part while retrieving and reading documents. Reading documents takes much longer and highlighting and writing the summary proposed in section 6.3 does not account for any significant extra time.

7.4 Stage 4: knowledge extraction

At this key stage, 25 concepts were defined using the types defined in Table 2 (see Table 5).

An Excel table was used to annotate documents when they met specific criteria, according to Table 5. A part of this work could be done when reading and highlighting documents. To complete this annotation task, the free program DocFetcher was used. Its outcome is a list of the files that meet the search criteria, showing the number of matches in each file, the context paragraph where the keywords were found, and a direct link to the files. These features make it possible to review any concept presence in 5-10 minutes when all the documents have been read, and it becomes extremely easy to carry out efficient searches. It is necessary to reject documents whose matches belong only to the "References" section. The total time dedicated to the 25 concepts defined was less than 4 hours. The outcome of this step is a table with the list of documents, their tags, summary, and concepts (Figure 7).

Figure 7 shows the concept map where most of the values are "x", there are values for precision and recall concepts, and there are names. The bottom line displays the count for the number of documents that meet each concept requirement. The use of the relative importance index (RII) method assigns distinct importance to the hits obtained in each document. This way, a weighted count is obtained for each concept. "Semantics" is the most important concept and is the basis for calculating the RII in every other concept. In this case "semantics" is a sort of wide concept because almost every document talks about semantics without a specific purpose, but that is not a problem as is shown in the next section.

Table 5 Key concepts for knowledge extraction.

Concept	Type	Explanation
Scientific papers	Condition	The document addresses scientific papers
IE	Keyword	Information extraction is considered
IR	Keyword	Information retrieval is considered
Improvement	Idea	Need for improvement of current IE/IR techniques
Concepts	Keyword	Concept as an entity, related to semantics and ontologies
Cosine	Keyword	Algorithm intended to evaluate the similarity
NLP	Keyword	Natural language process is cited
Knowledge	Keyword	Knowledge extraction concept is cited
ANN	Keyword	Artificial neural network is cited
Fuzzy	Keyword	Fuzzy techniques and fuzzy logic are cited
Bayes	Keyword	Bayes decision function (classification method) is cited
Semantics	Keyword	Semantics is cited
Ontology	Keyword	Ontology is cited
Query	Keyword	Query is cited, usually related to Boolean operations
Rule-based	Keyword	Rule-based and rule are cited related to queries
Clustering	Keyword	Clustering technique is used to classify documents
Machine learning	Keyword	Machine learning is cited
Artificial intelligence	Keyword	Artificial intelligence is cited
Manual	Idea	Manual operation is needed for supervision, training, etc.
System	Keyword	A system is proposed, although different in each paper
Precision	Numeric	Percentage of precision yielded by the proposed system
Recall	Numeric	Percentage of recall yielded by the proposed system
Tags	Keyword	Administrative tags are used and retrieved
Specific activity	Name	The document addresses some specific kind of papers
Specific country	Name	The document addresses some specific country

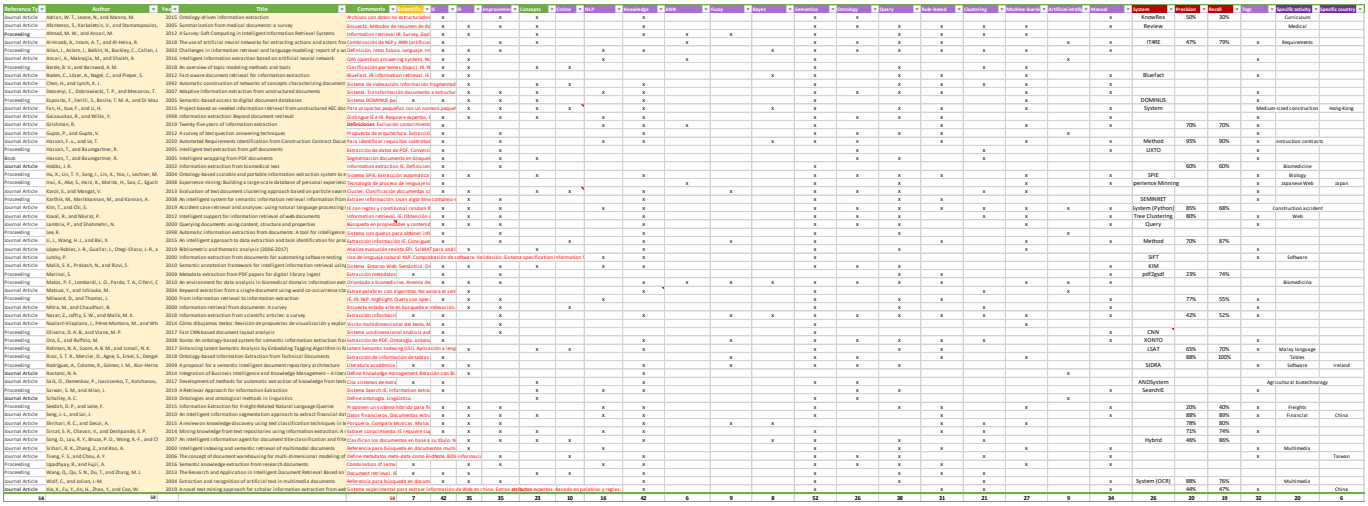

Figure 7 Reference list with concepts.

8. RESULTS AND DISCUSSION

The results of the knowledge extraction performed according to the proposed method can be expressed by using the concepts defined and their RII. A ranked list of concepts using the RII gives an accurate view of how scientists address information extraction as a gate to knowledge extraction (Table 6) and a Pareto diagram gives a better understanding of the relative importance of each concept (Figure 8).

Table 6 Ranked list of concepts.

#	Concept	RII
1	Semantics	100%
2	Knowledge	81%
3	IE	78%
4	Query	74%
5	Improvement	69%
6	IR	69%
7	Manual	66%
8	Tags	63%
9	Rule-based	61%
10	Machine learning	55%
11	Ontology	49%
12	Concepts	47%
13	Clustering	45%
14	System	44%
15	Precision	40%
16	Recall	38%
17	Specific activity	33%
18	NLP	30%
19	Cosine	23%
20	Fuzzy	17%
21	Artificial intelligence	17%
22	Bayes	14%
23	Scientific papers	12%
24	Specific country	11%
25	ANN	11%

It is remarkable that "knowledge extraction" is the second most cited concept, after "semantics," whose presence is compulsory in this kind of documents. "Information extraction" is placed third in the list and "information retrieval" is sixth, although the search string was "intelligent information extraction". This proves how close both concepts are in the literature.

Figure 8 proves that the results obtained do not follow the Pareto rule. It is possible to differentiate three groups according to concept relevance: 1 to 9, 10 to 18, and 18 to 25.

The first group includes basic concepts related to automatization, e.g., "query" and "rule-based". However, this group contains concepts indicating that there are strong limitations in the state of the art: "Need for improvement of current IE/IR techniques" is placed fifth and "Manual operation is needed for supervision, training, etc." is placed seventh. "Tags" is placed eighth (administrative tags) and this fact proves that the solutions proposed to extract information frequently address tags, less relevant than insights information.

The second group includes concepts related to the technology applied to retrieve and extract information (machine learning, ontologies, concepts, and clustering). It also includes the concept "system" that represents all the systems proposed. All of them are different and, for that reason, they were grouped in that concept to make it possible to give them some visibility. The concept "specific activity," placed seventeenth, shows that a significant part of the documents studied are intended for a specific purpose, and that fact makes them less applicable to this study. This group includes the concepts "precision" and

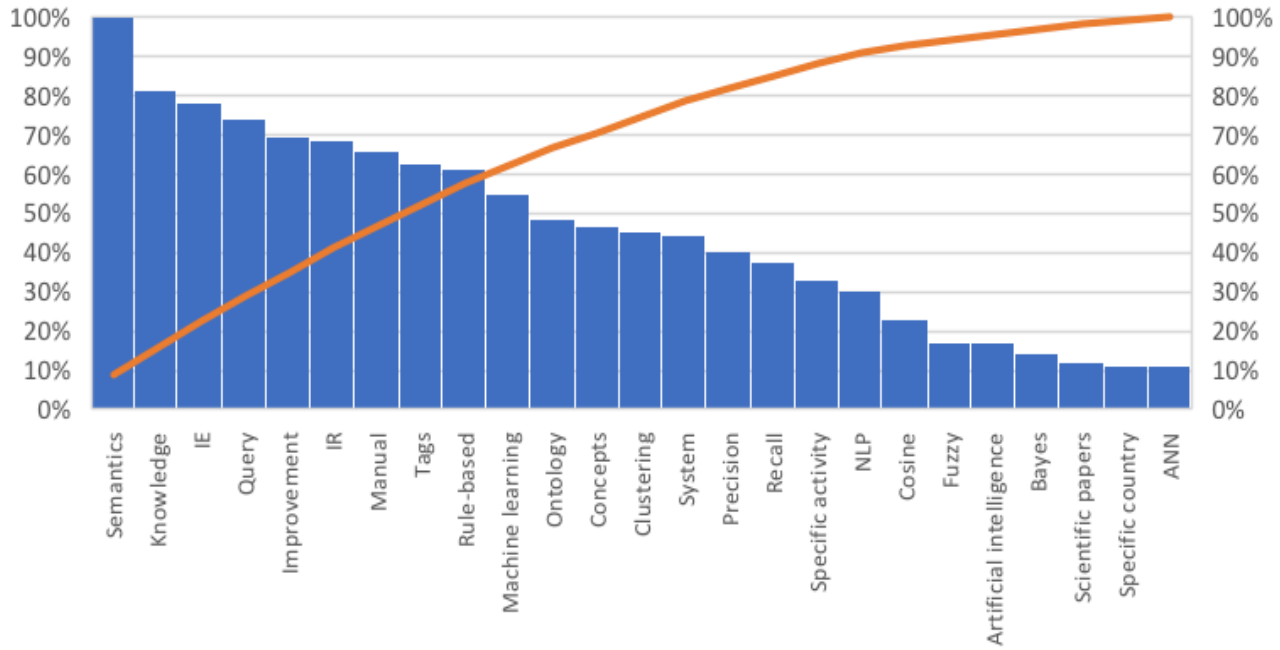

Figure 8 Pareto diagram of concepts using their RII.

"recall": the average values for precision and recall in the literature review performed are 64% and 70%, respectively, which are very far from a comfortable confidence level.

The third group contains the least relevant concepts and they are related to the most sophisticated techniques, e.g., "artificial intelligence." This seems to prove that they are far from a mature state that would allow them to be commonplace. The concept "scientific papers" is placed twenty-third because only seven out of the 58 documents studied address this subject.

The specific field of knowledge extraction from scholarly documents asks for affordable solutions that are easy to work with. Nassar says that "Manual analysis is not scalable and efficient" and cites other authors who state that a systematic literature review could take 1 to 3 years (Nasar et al. 2018). This study has used a manual method to extract knowledge starting with a systematic literature review, and the whole process took less than one month. The results presented in this study prove that knowledge extraction can be efficiently performed manually with the help of desktop tools that are commonplace. It does not matter that manual analysis is not scalable because researchers usually face a scholarly library with only a few hundred documents in each research project. The method proposed was also used in a distinct research project with a library that held 300 documents (Vegas-Fernández 2019). In practice, document reading takes up most of the time dedicated to literature review in a research project, much more than retrieving and organizing documents. This paper proposes a feasible way to optimize knowledge extraction, giving up, for now, the option of a fully automatic information retrieval and extraction system, and proposing "concept definition" as the most relevant task.

9. CONCLUSIONS

Technique algorithms are not always the answer to efficient extraction of information from scholarly document databases and sophisticated automatic systems do not seem to be the best fit to solve the researcher's needs. Any possible automated solution that requires manual training, supervision, and tuning is not worthwhile because it requires too much time dedicated to those tasks and it is shorter and more efficient to do it by hand.

The relevance of concept definition has frequently been underestimated and this paper proposes and proves that proper concept definition is key to achieve outstanding knowledge extraction. The results of the analysis conducted with a scholarly document database confirm the suitability of the approach and the method that has been explained.

This paper has presented a simple but efficient method that takes advantage of free desktop tools that are commonplace. By following this method, it is very easy to carry out a systematic literature review, in order to

retrieve, filter, and organize results, and to extract information to transform it into knowledge. The conceptual basis is a semantics-oriented concept definition and a relative importance index to measure concept relevance in the literature studied.

The detailed explanation of the proposed procedure in four steps shows that most of the tasks require mental activity that cannot be helped by automated systems.

The method proposed is intended for knowledge extraction from scholarly document databases, but it could also be used in other projects such as departmental document databases whenever the total number of documents in the library is only a few hundred.

10. REFERENCES

Adrian, W. T., Leone, N., and Manna, M. (2015). "Ontology-driven information extraction." *arXiv preprint arXiv:1512.06034*.

Afantenos, S., Karkaletsis, V., and Stamatopoulos, P. (2005). "Summarization from medical documents: a survey." *Artificial intelligence in medicine*, 33(2), 157-177.

Ahmad, M. W., and Ansari, M. "A survey: soft computing in intelligent information retrieval systems." *Proc., 2012 12th International Conference on Computational Science and Its Applications*, IEEE, 26-34.

Al-Hroob, A., Imam, A. T., and Al-Heisa, R. (2018). "The use of artificial neural networks for extracting actions and actors from requirements document." *Information and Software Technology*, 101(2018), 1-15.

Alashwal, A. M., and Al-Sabahi, M. H. (2018). "Risk factors in construction projects during unrest period in Yemen." *Journal of Construction in Developing Countries*, 23(2), 43–62.

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., and Harper, D. J. "Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval." *Proc., ACM SIGIR Forum*, ACM New York, NY, USA, 31-47.

Ansari, A., Maknojia, M., and Shaikh, A. (2016). "Intelligent information extraction based on artificial neural network." *International Journal in Foundations of Computer Science & Technology*, 6(1).

Barde, B. V., and Bainwad, A. M. (2018). "An overview of topic modeling methods and tools." *Proc., 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 745-750.

Bettany-Saltikov, J. (2012). *How to do a systematic literature review in nursing: a step-by-step guide*, McGraw-Hill Education (UK), Maidenhead, UK.

Boden, C., Löser, A., Nagel, C., and Pieper, S. (2012). "Fact-aware document retrieval for information extraction." *Datenbank-Spektrum*, 12(2), 89-100.

Buzan, T. (2004). *Cómo crear mapas mentales*, Ediciones Urano, Barcelona, Spain.

Chen, H., and Lynch, K. J. (1992). "Automatic construction of networks of concepts characterizing document databases." *Ieee T Syst Man Cyb*, 22(5), 885-902.

Dezsenyi, C., Dobrowiecki, T. P., and Meszaros, T. (2007). "Adaptive information extraction from unstructured documents." *International Journal of Intelligent Information and Database Systems*, 1(2), 156-180.

Esposito, F., Ferilli, S., Basile, T. M. A., and Di Mauro, N. (2005). "Semantic-based access to digital document databases." *Proc., International Symposium on Methodologies for Intelligent Systems*, Springer, Berlin, Heidelberg, Germany, 373-381.

Fan, H., Xue, F., and Li, H. (2015). "Project-based as-needed information retrieval from unstructured AEC documents." *Journal of Management in Engineering*, 31(1), A4014012.

Gaizauskas, R., and Wilks, Y. (1998). "Information extraction: Beyond document retrieval." *Journal of documentation*, 54(1), 70-105.

Grishman, R. (2019). "Twenty-five years of information extraction." *Natural Language Engineering*, 25(6), 677-692.

Gupta, P., and Gupta, V. (2012). "A survey of text question answering techniques." *International Journal of Computer Applications*, 53(4), 1–8.

Hassan, F. u., and Le, T. (2020). "Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 12(2), 04520009.

Hassan, T., and Baumgartner, R. "Intelligent text extraction from pdf documents." *Proc., International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, IEEE, 2–6.

Hassan, T., and Baumgartner, R. (2005b). *Intelligent wrapping from PDF documents*, CEUR Workshop Proceedings, Točná, Czech Republic.

Hobbs, J. R. (2002). "Information extraction from biomedical text." *Journal of biomedical informatics*, 35(4), 260-264.

Hu, X., Lin, T. Y., Song, I., Lin, X., Yoo, I., Lechner, M., and Song, M. "Ontology-based scalable and portable information extraction system to extract biological knowledge from huge collection of biomedical web documents." *Proc., IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, IEEE, 77-83.

Inui, K., Abe, S., Hara, K., Morita, H., Sao, C., Eguchi, M., Sumida, A., Murakami, K., and Matsuyoshi, S. "Experience mining: Building a large-scale database of personal experiences and opinions from web documents." *Proc., 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, 314-321.

Jarkas, A. M., and Haupt, T. C. (2015). "Major construction risk factors considered by general contractors in Qatar." *Journal of Engineering, Design and Technology*, 13(1), 165–194.

Karol, S., and Mangat, V. (2013). "Evaluation of text document clustering approach based on particle swarm optimization." *Open Computer Science*, 3(2), 69-90.

Karthik, M., Marikkannan, M., and Kannan, A. "An intelligent system for semantic information retrieval information from textual web documents." *Proc., International Workshop on Computational Forensics*, Springer, Berlin, Heidelberg, Germany, 135-146.

Kasperiuniene, J., and Zydziunaite, V. (2019). "A systematic literature review on professional identity construction in social media." *SAGE Open*, 9(1), 2158244019828847.

Kim, T., and Chi, S. (2019). "Accident case retrieval and analyses: using natural language processing in the construction industry." *Journal of Construction Engineering and Management*, 145(3), 04019004.

Koval, R., and Návrat, P. (2012). "Intelligent support for information retrieval of web documents." *Computing and Informatics*, 21(5), 509–528.

Lambrix, P., and Shahmehri, N. (2000). "Querying documents using content, structure and properties." *Journal of Intelligent Information Systems*, 15(3), 287-307.

Lee, R. "Automatic information extraction from documents: A tool for intelligence and law enforcement analysts." *Proc., Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*, AAAI Press Menlo Park, CA.

Li, J., Wang, H. J., and Bai, X. (2015). "An intelligent approach to data extraction and task identification for process mining." *Information Systems Frontiers*, 17(6), 1195-1208.

López-Robles, J.-R., Guallar, J., Otegi-Olaso, J.-R., and Gamboa-Rosales, N.-K. (2019). "Bibliometric and thematic analysis (2006-2017)." *El profesional de la información*, 28(4), e280417.

Lutsky, P. (2000). "Information extraction from documents for automating software testing." *Artificial Intelligence in Engineering*, 14(1), 63-69.

Malik, S. K., Prakash, N., and Rizvi, S. (2010). "Semantic annotation framework for intelligent information retrieval using KIM architecture." *International Journal of Web & Semantic Technology (IJWest)*, 1(4), 12-26.

Marinai, S. "Metadata extraction from PDF papers for digital library ingest." *Proc., 2009 10th International conference on document analysis and recognition*, IEEE, 251-255.

Matos, P. F., Lombardi, L. O., Pardo, T. A., Ciferri, C. D., Vieira, M. T., and Ciferri, R. R. (2010). "An environment for data analysis in biomedical domain: information extraction for decision support systems." *Proc., International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, Berlin, Heidelberg, Germany, 306-316.

Matsuo, Y., and Ishizuka, M. (2004). "Keyword extraction from a single document using word

co-occurrence statistical information." *International Journal on Artificial Intelligence Tools*, 13(01), 157-169.

Milward, D., and Thomas, J. "From information retrieval to information extraction." *Proc., ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 85-97.

Mitra, M., and Chaudhuri, B. (2000). "Information retrieval from documents: A survey." *Information retrieval*, 2(2-3), 141-163.

Nagalla, V., Dendukuri, S. C., and Asadi, S. S. (2018). "Analysis of risk assessment in construction of highway projects using relative importance index method." *International Journal of Mechanical Engineering and Technology*, 9(3), 1–6.

Nasar, Z., Jaffry, S. W., and Malik, M. K. (2018). "Information extraction from scientific articles: a survey." *Scientometrics*, 117(3), 1931-1990.

Nualart-Vilaplana, J., Pérez-Montoro, M., and Whitelaw, M. (2014). "Cómo dibujamos textos: Revisión de propuestas de visualización y exploración textual." *El profesional de la información*, 23(3), 221-235.

Oliveira, D. A. B., and Viana, M. P. (2018). "Fast CNN-based document layout analysis." *Proc., Proceedings of the IEEE International Conference on Computer Vision Workshops*, IEEE Computer Society, 1173-1180.

Oro, E., and Ruffolo, M. "Xonto: An ontology-based system for semantic information extraction from pdf documents." *Proc., 2008 20th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, 118-125.

Rahman, N. A., Soom, A. B. M., and Ismail, N. K. "Enhancing Latent Semantic Analysis by Embedding Tagging Algorithm in Retrieving Malay Text Documents." *Proc., Asian Conference on Intelligent Information and Database Systems*, Springer, 309-319.

Renault, B. Y., and Agumba, J. N. (2016). "Risk management in the construction industry: a new literature review." *MATEC Web of Conferences*, 66(2016), 0008.

Rizvi, S. T. R., Mercier, D., Agne, S., Erkel, S., Dengel, A., and Ahmed, S. (2018). "Ontology-based Information Extraction from Technical Documents." *Proc., ICAART (2)*, Science and Technology Publications, Lda, 493-500.

Rodríguez, A., Colomo, R., Gómez, J. M., Alor-Hernandez, G., Posada-Gomez, R., Juarez-Martinez, U., Gayo, J. E. L., and Vidyasankar, K. "A proposal for a semantic intelligent document repository architecture." *Proc., 2009 Electronics, Robotics and Automotive Mechanics Conference (CERMA)*, IEEE, 69-75.

Rostami, A., Sommerville, J., Wong, I. L., and Lee, C. (2015). "Risk management implementation in small and medium enterprises in the UK construction industry." *Engineering, Construction and Architectural Management*, 22(1), 91–107.

Saik, O., Demenkov, P., Ivanisenko, T., Kolchanov, N., and Ivanisenko, V. (2017). "Development of methods for automatic extraction of knowledge from texts of scientific publications for the creation of a knowledge base Solanum TUBEROSUM." *Agricultural Biology*, 52(1), 1.

Sarwar, S. M., and Allan, J. "A Retrieval Approach for Information Extraction." *Proc., Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, Association for Computing Machinery, 249-252.

Schalley, A. C. (2019). "Ontologies and ontological methods in linguistics." *Language and Linguistics Compass*, 13(11), e12356.

Seedah, D. P., and Leite, F. (2015). "Information Extraction for Freight-Related Natural Language Queries." *Proc., Computing in Civil Engineering 2015*, American Society of Civil Engineers, 427-435.

Seng, J.-L., and Lai, J. (2010). "An Intelligent information segmentation approach to extract financial data for business valuation." *Expert Systems with Applications*, 37(9), 6515-6530.

Shrihari, R. C., and Desai, A. (2015). "A review on knowledge discovery using text classification techniques in text mining." *International Journal of Computer Applications*, 111(6).

Sirsat, S. R., Chavan, V., and Deshpande, S. P. (2014). "Mining knowledge from text repositories using information extraction: A review." *Sadhana-Acad P Eng S*, 39(1), 53-62.

Snyder, H. (2019). "Literature review as a research methodology: An overview and guidelines." *Journal of Business Research*, 104(2019), 333–339.

Song, D., Lau, R. Y., Bruza, P. D., Wong, K.-F., and Chen, D.-Y. (2007). "An intelligent

information agent for document title classification and filtering in document-intensive domains." *Decision Support Systems*, 44(1), 251-265.

Srihari, R. K., Zhang, Z., and Rao, A. (2000). "Intelligent indexing and semantic retrieval of multimodal documents." *Information Retrieval*, 2(2-3), 245-275.

Tseng, F. S., and Chou, A. Y. (2006). "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence." *Decision Support Systems*, 42(2), 727-744.

Upadhyay, R., and Fujii, A. "Semantic knowledge extraction from research documents." *Proc., 2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, 439–445.

Vegas-Fernández, F. (2019). "Factor de visibilidad. Nuevo indicador para la evaluación cuantitativa de riesgos." PhD PhD, Universidad Politécnica de Madrid, Universidad Politécnica de Madrid.

Vegas-Fernández, F., and Rodríguez López, F. (2019). "Risk management improvement drivers for effective risk-based decision-making." *Journal of Business, Economics and Finance (JBEF)*, 8(4), 223–234.

Wang, Q., Qu, S. N., Du, T., and Zhang, M. J. "The Research and Application in Intelligent Document Retrieval Based on Text Quantification and Subject Mapping." *Proc., Advanced Materials Research*, Trans Tech Publ, 2561-2568.

Wolf, C., and Jolion, J.-M. (2004). "Extraction and recognition of artificial text in multimedia documents." *Formal Pattern Analysis & Applications*, 6(4), 309-326.

Xia, N., Zou, P. X., Griffin, M. A., Wang, X., and Zhong, R. (2018). "Towards integrating construction risk management and stakeholder management: A systematic literature review and future research agendas." *International Journal of Project Management*, 36(5), 701–715.

Xie, X., Fu, Y., Jin, H., Zhao, Y., and Cao, W. (2019). "A novel text mining approach for scholar information extraction from web content in Chinese." *Future Generation Computer Systems*.