Automatic Generation of Links in Patent Documents

C. M. Souza, M. E. Santos, M. R. G. Meireles

Pontifical Catholic University of Minas Gerais, Brazil
cinthia.mikaela@sga.pucminas.br
matheus.santos.1004060@sga.pucminas.br
magali@pucminas.br

Abstract. Patents are organized into classification systems, which assist offices and users in the process of seeking and retrieving such documents. A wide variety of users use the patent systems and the information contained in these documents. In addition to office professionals, such as examiners and analysts, who determine whether the invention satifies the conditions required to be patented and summarize the content of the document, other users such as inventors, researchers, investors and business managers have a keen interest in understanding the content of patents.However, patents are complex legal documents with a significant number of technical and descriptive details, which makes it difficult to identify and analyze the information contained in these documents. An automatic link system associated with some of the terms found in the patents would provide quick access to the concepts contained in specific knowledge bases. This work presents partial results of a project whose objective is the automatic generation of links in patent documents. The experiments were conducted with four subgroups of the United States Patent and Trademark Office (USPTO), which uses the Cooperative Patent Classification (CPC) classification system. In a first step, since documents do not have keywords, meaningful terms were selected to be designated as link origins, using the algorithm X^2.Once the link destinies were selected, in a later step, keywords with more than one meaning were disambiguated. It is expected, with the creation of automated links, to aid in the reading of patent texts, thus making it easier to access concepts related to the terms presented by the documents and to the understanding of the information disclosed by the inventors.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords: Disambiguation, Keywords Extraction, Link creation, Patents

1. INTRODUCTION

Patents are an important knowledge source and, therefore, their analysis has been considered a useful tool for research and for management development. Ouellette [2017] conducted a survey of 832 researchers to assess the importance of patent study. Most researchers in different fields of knowledge have stated that they have found useful information in the documents, but acknowledged that there is room for improvement, particularly as regards the accessibility and understanding of information contained in patents. Many of the interviewees stated that it is possible to find information unavailable in the scientific literature and that patents are an underutilized complement to the dissemination of scientific knowledge.

In order to extract knowledge of the information contained in the patents, it is necessary to deal with the difficulty of understanding the texts, which are, in particular, complex, with technological details, legal language and exhaustive descriptions [Meireles et al. 2016]. In this context, conventional approaches to information retrieval are difficult to apply and therefore the in-depth study of patents and their consequences has yet to become more accessible, identifying potential areas of research for the scientific community and generating useful information in processes of decision-making in the

area of competitive intelligence. The proposal to create an automatic link generation system, which allows simplified access to the concepts related to the terms presented by patents, meets this problem, beckoning with the possibility of quickly accessing the knowledge base related to the theme proposed by the patent.

The automatic determination of a link includes the identification of possible text fragments that should be associated with knowledge bases. In most cases, keywords are selected and their extraction can be done by supervised or unsupervised methods [Mihalcea and Csomai 2007]. Before selecting the texts that will be associated with certain words, it is necessary to define the context in which the word is being used. This is because several language units have different meanings, which is a common feature in many languages and a problem that needs to be addressed in depth.

This paper presents partial results of a project that aims at the automatic generation of links in patent documents. The algorithm used for keyword extraction was the X^2, defined in the work of Reginaldo et al. [2017], as the one that achieved better results in this context. The work was divided into 5 sections. Section 2 presents the main concepts used, as well as a description of the algorithms implemented in the processes of keyword extraction of links and disambiguation. Section 3 presents the proposed methodology, which discuss the database used and the methodological steps of the project. Sections 4 and 5 show the results, analyzes and final considerations.

2. AUTOMATIC LINK GENERATION

The automatic link generation process can be divided into two distinct steps, identifying the source of a link and determining the appropriate document to be associated with the terms selected as the source. In the second step, we must solve the problem of disambiguation of the meaning of the word or term. Given this, this section will be divided into four subsections that will address the processes of keyword extraction of links and disambiguation.

2.1 Keywords Extraction

In some cases, the keywords are not defined by the authors of the document and therefore it is necessary to develop a method or select an algorithm that extracts the words considered significant for the document and that can represent it in a system of recovery of information.

In this work, the algorithm used for the extraction of keywords was the X^2, used by Mihalcea and Csomai [2007] . This algorithm evaluates the independence between two variables and compares observed and expected values, evaluating how far apart they are. This algorithm is used to order the words according to their dependence on the patent, so that the greater the note given by X^2 to a word, the greater its dependence on the document. Even if the algorithm accepts that a word is independent, the note given by it to the word is simply added to the ordering in lower positions [Reginaldo et al. 2017].

2.2 Identification of the Link's Destination

To identify the destination of the link, an approach similar to that presented by Jana et al. [2017] was used. For this, we implemented an algorithm that uses the wikipedia python library, which encapsulates the MediaWiki API[1] for this purpose. This library allows access to Wikipedia data and metadata via API. In order to access this data, the user provides an input data and the algorithm provides a Wikipedia page. In some cases, the attribute provided may have ambiguous meaning. If this occurs, the algorithm generates an exception, stating the need to treat it. The algorithm sends the keyword with the meaning obtained, using the disambiguation algorithm, described in the next

[1]Application Programming Interface

subsection, and receives the url of the specified page. For the effective creation of the link in the patent document, the keyword is replaced by the url in the HTML code of the page.

2.3 Disambiguation of the Word's Meaning

The techniques of disambiguation aim to computationally identify the meaning of a word, taking into account the context in which it is inserted. Therefore, given a document with a sequence of words $T = w_1, w_2, ..., w_n$, this technique aims to give meaning to all or some words of that document. This task can be performed with only one lexical sample or with all the words in the document. Generally, the lexical sample is more used, because a wide coverage of domains is necessary to carry out the disambiguation of all the words [Corrêa Jr et al. 2018].

In this work, a disambiguation algorithm proposed by Panchenko et al. [2017] was used. This algorithm receives, as input, the word that will be disambiguated and the context to which it belongs. In addition, it is necessary to define some parameters, such as the model used and the output format. Among the available models, the ensemble model is the most complete, since it searches for the meaning of the ambiguous word in the inventory of word meanings. If a word is outside this vocabulary, then it is disambiguated using the super meaning inventory. This template was created from a text corpus which is a combination of Wikipedia, ukWaC, corpus LCC News and Gigaword. For the realization of disambiguation, this algorithm performs the following steps:

—Extraction of context features computing word and feature similarities;
—Word meaning induction;
—Labeling of clusters with hypernyms and images (hypernym is a word with a broad meaning constituting a category which words with more specific meanings fall into[2]);
—Disambiguation of words in context based on the induced inventory.

In the end, the algorithm returns the meaning of the word, its hypernyms, the set of related words taken from the dictionary, a set of phrases to exemplify the meaning of the disambiguated word. The context words are those that co-occur with the word ambiguous destination in the given meaning, and they are also returned with the words related to the disambiguated word, taken from the text itself, and the level of trust of the disambiguation. The confidence level is a metric that evaluates the disambiguation result. It is calculated from the extraction of hypernyms. For this, the algorithm ranks the hypernyms using functions that relate the word to the set of words of the cluster and to the hypernym.

2.4 Related Works

Linking web data to relevant knowledge base articles has become popular, and because of this, some research on automatic linking of text to important knowledge base articles has captured the interest of the research community [Gardner and Xiong 2009]. The majority of the work is focused in linking Wikipedia texts to their referent Wikipedia pages [Mihalcea and Csomai 2007; Cucerzan 2007; Jana et al. 2017]

Mihalcea and Csomai [2007] used Wikipedia for automatic extraction of keywords and for the disambiguation process. The system developed by the authors automatically extracts the keywords, makes the disambiguation process, and generates the link with the Wikipedia page. For the extraction of the keyword, three methods were tested, tf-idf, X_2, and Keyphraseness. For the disambiguation of the keyword, the authors tested two methods, the first one, a knowledge-based approach, and another

[2]en.oxforddictionaries.com

one, based on data-driven. In the end, the Wikify! system presented superior results relating to the competitive baselines.

Jana et al. [2017] presented a project to generate links in abstracts of scientific documents with Wikipedia articles. They performed the extraction of the important mentions of the scientific text using tf-idf, together with a set of intelligent filters. Afterwards, for each mention, they extracted a list of candidate entities (Wikipedia links). These entities were classified and punctuated according to the similarity, and finally, based on this score, the entity for link generation was selected. The results show that the methodology used helps to improve the performance of the wikification task in scientific articles

3. METHODOLOGY

3.1 Database

The database used in the experiment is provided by the United States Patent and Trademark Office (USPTO), whose classification system is the Cooperative Patent Classification (CPC). CPC classifies patents into sections, classes, subclasses, groups, and subgroups. For this work, four subgroups, G06K7/1443, G06K7/1447, G06K7/1452 and G06K7/1456 of the G06K subclass, named recognition of data, presentation of data, record carriers, handling record carriers, were selected. Figure 1 illustrates the organization of this patent database.

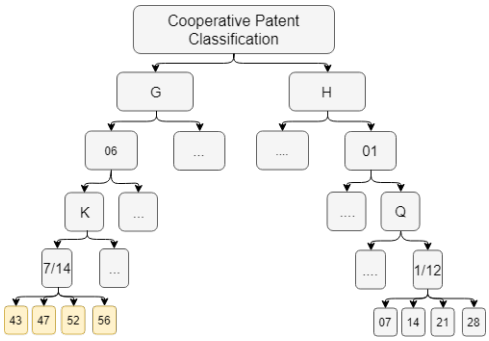

Fig. 1. Organization of the Database

The database used is composed of 910 patents and was updated on June 12, 2018. To validate the methodology in this work, 10 patents of each subgroup were selected. Table I shows the name of the subgroup in the CPC classification system and the distribution of the patents in each of them.

Table I. Database

Code CPC	Number of patents
G06K 7/1443	452
G06K 7/1447	263
G06K 7/1452	78
G06K 7/1456	117

3.2 Methodological Steps

In the first step, the patent summaries were pre-processed. The algorithm used performs the removal of stopwords, special characters and the stemization of words. In addition, the algorithm uses a vocabulary based on Wikipedia titles to generate significant n-grams.

After the preprocessing has been executed, there is an array of documents by words where the occurrence of the words in the document are computed. The second step consists of extracting keywords using the algorithm X^2. This algorithm receives as input the matrix generated by the preprocessing and returns a list of keywords.

In the third step, the preparation of the output data from the algorithm X^2 for the input of the link generation algorithm was performed. For this, an algorithm that performs the search for the keywords extracted by X^2 and extracts those that have the three highest values generated by the algorithm was implemented. For each patent, three keywords were obtained. After this phase, the paragraphs containing the words for the entry of the disambiguation algorithm are prepared, with the removal of blanks and invalid characters.

In the fourth step, the link destination was identified. For this, an algorithm was implemented in python, using the "wikipedia"library. This algorithm receives a keyword and searches a Wikipedia page with the corresponding content and returns the page link. After this phase, the keyword is replaced by the link in the HTML code of the page. However, in some cases, ambiguity of the meaning of the keyword may occur. In this case, the algorithm throws an exception, notifying that there is more than one content-related page.

For the treatment of the exception, a disambiguation algorithm was used. This algorithm receives the keyword and a paragraph from the patent text that owns the keyword. In this case, this paragraph is the context that the algorithm will use to find the meaning of the word. For this work, the ensemble model was chosen. This algorithm returns the meaning of the word and four hypernyms. The reliability of the disambiguation process is then verified with the use of the metric provided by the algorithm. The trust value (0 to 100%) indicates whether the meaning of the keyword and the content of the page selected for the link's destination are associated with the same context. Figure 2 presents a diagram, exemplifying each step of the proposed methodology.

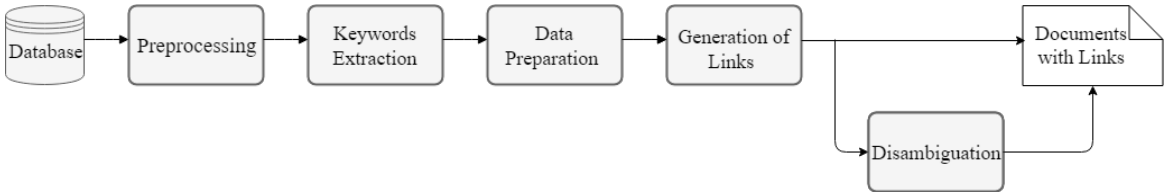

Fig. 2. Proposed methodology

4. RESULTS AND DISCUSSION

Initially, the documents were preprocessed to generate the input to the X^2 algorithm. Three keywords were extracted from each patent document. The next algorithm receives the keyword and finds the destination of the link. In some cases, the keyword may have ambiguous meaning. In this case, the meaning of the word is disambiguated. Tables II, III, IV and V present some of the results obtained in the disambiguation stage. The first column presents the extracted keywords that have ambiguous meaning, the second one, the hypernyms, the third one, the meaning of the word in the specific context of the patent and the fourth presents the reliability calculated by the algorithm of disambiguation. Before each table, the three extracted keywords and the context for disambiguation of the selected keywords are presented. Patents were identified as P_1, P_2, P_3 and P_4.

Keywords of patent P_1: dibit, data strip, buffer.

Context for disambiguation of the keyword buffer:

"An optical reader for reading high density dibit encoded data from a data strip comprises an optical detector (13) and a detector (14,15,DSP) connected to the optical detector (13) and arranged to decode its output. The detector includes a row of photosensitive elements arranged to extend transversely across a row of dibits and to form an image of the row of dibits. The decoder (14,15,DSP) includes a buffer (15) arranged to store the row image. An image transform device detects a skew angle of the strip and transforms the stored image to compensate for the distortion of the image."

Table II. Results of the disambiguation process for P_1

Keyword	Hypernyms	Meaning	Reliability
buffer	feature factor thing	type	83,72%

Keyword of patent P_2: barcode, components, bitmap.

Context for disambiguation of the keyword components:

"A technique for reading a bar code representative of message information is stored electronically in bit-map form. The bit map is obtained by optically scanning a document carrying non-bar code information also to convert pixel information into bit information. A row is identified in the bit map, which is the most likely one to pass through an area in the bit map containing bar code information. The data stored in the bit map is compared with assigned data corresponding to designated bar code components. Once such a row is identified, processing of rows above and below it in the bit map reveals whether rows more likely to agree with the coded message information can be found. Initially, a fast processing mode is selected involving processing of only a few rows to maximize processing speed in case acceptable data can be found. If acceptable bar code data is not found, then the processing is switched into a slow processing mode, which involves processing of a greater number of rows. Once a row likely to contain the bar code data of interest is identified, the information it contains is converted into signals, which are used for converting the coded bar code data into the message represented thereby."

Table III. Results of the disambiguation process for P_2

Keyword	Hypernyms	Meaning	Reliability
components	area product company application	service	100%

Keywords of patent P_3: characteristics, magnetic, material magnetic.

Context for disambiguation of the keyword characteristics:

"Appartus and methods of verifying objects utilize detectable characteristics of a plurality of spaced apart, magnetizable magnetic security regions affixed to the object. Magnetic characteristics for each magnetic region are detected from two different orientations. A verification apparatus senses the magnetic characteristics from two different orientations and compares it to a prestored representative profile previously created. Correspondence between the prestored profile and the currently read characteristics indicates an authentic object. Other objects can be rejected."

Table IV. Results of the disambiguation process for P_3

Keyword	Hypernyms	Meaning	Reliability
characteristics	area topic service factor	issue	100%

Keyword of patent P_4: function code, link, data object.

Context for disambiguation of the keyword link:

"A given data object can effectively contain both a graphical representation to a network user and embedded information, such as the URL address of another network node, thereby to permit the object itself to serve as an automated hot link. The underlying development tools and web site browsers create and identify such an object for use in a manner similar to a hot link, as provided on the World Wide Web."

Table V. Results of the disambiguation process for P_4

Keyword	Hypernyms	Meaning	Reliability
link	factor feature thing area	issue	87,82%

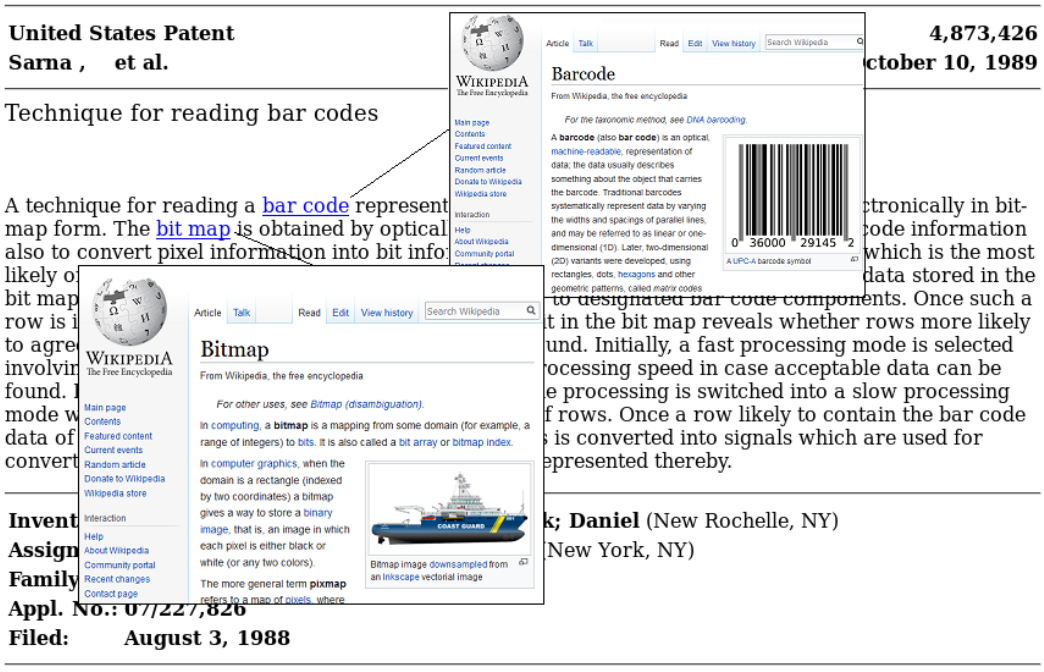

Fig. 3. Results for patent P_2

Figure 3 exemplifies the end result, showing the links in the patent and their destination. By analyzing the presented results, it can be seen that the reliability values found by the disambiguation

algorithm were satisfactory. However, the algorithm did not perform well with some keywords, which needs to be investigated. It is worth mentioning that the algorithm of link generation was able to identify, in most cases, that the keyword had more than one meaning and, therefore, it was necessary to treat this problem using the disambiguation algorithm. However, in some cases, it was not possible to find a Wikipedia base page with the keyword in the meaning provided, and, in these situations, the texts of other patents could be incorporated into the knowledge base.

5. FINAL CONSIDERATIONS

Patents are legal documents with a significant number of technical and descriptive details, which makes their analysis very complex. Access to information in such documents is often laborious, due to the difficulty imposed by technical language and poorly designed writing styles, contrary to the main objective of a patent system of sharing knowledge.

The proposal to create an automatic link generation system, which allows simplified access to the concepts related to the terms presented by the patents, is an alternative to provide a simpler access to the knowledge bases related to the theme proposed by the patent. With this proposal, it is expected to contribute to the study of links in patents, facilitating the understanding of the information contained in these documents and promoting the dissemination of scientific knowledge associated with the technological advances proposed by these inventions. As a development of this work, a user experiment will be conducted to determine which terms in the patents used in the experiments require reference to external knowledge. As users with different types of background would indicate different words to link, these selected words can be considered as gold standard to evaluate the performance of our experiments.

6. ACKNOWLEDGEMENTS

The authors thank the financial support of the Pontifical Catholic University of Minas Gerais, the National Council for Scientific and Technological Development (CNPq, grant 429144/2016-4) and the Foundation for Research Support of the State of Minas Gerais (FAPEMIG, grant APQ 01454-17).

REFERENCES

Corrêa Jr, E. A., Lopes, A. A., and Amancio, D. R. Word sense disambiguation: A complex network approach. *Information Sciences* vol. 442–443, pp. 103–113, 2018.

Cucerzan, S. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. pp. 708–716, 2007.

Gardner, J. J. and Xiong, L. Automatic link detection: A sequence labeling approach. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. ACM, New York, NY, USA, pp. 1701–1704, 2009.

Jana, A., Mooriyath, S., Mukherjee, A., and Goyal, P. Wikim: metapaths based wikification of scientific abstracts. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1–10, 2017.

Meireles, M. R. G., Ferraro, G., and Geva, S. Classification and information management for patent collections: a literature review and some research questions. *Information Research* 21 (1), 2016.

Mihalcea, R. and Csomai, A. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. ACM, pp. 233–242, 2007.

Ouellette, L. L. Who reads patents? *Nature biotechnology* 35 (5): 421–424, 2017.

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., and Biemann, C. Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 1. pp. 86–98, 2017.

Reginaldo, T. V., Lucindo, D. L. B., Meireles, M. R. G., Patrocínio Júnior, Z. K. G., and Almeida, P. E. M. A comparison of algorithms for the extraction of keywords in a patent database. *Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering*, 2017.