A COMPARISON OF ALGORITHMS FOR THE EXTRACTION OF KEYWORDS IN A PATENT DATABASE

Thiago V. Reginaldo

Daniel L. B. Lucindo

Magali R. G. Meireles

Zenilton K. G. do Patrocínio Júnior

thiagopesquisa42@gmail.com

daniellblucindo@gmail.com

magali@pucminas.br

zenilton@pucminas.br

Instituto de Ciências Exatas e Informática, Pontifícia Universidade Católica de Minas Gerais

Rua Walter Ianni, 255, São Gabriel, Belo Horizonte, 31.980-110, Minas Gerais, Brazil

Paulo E. M. de Almeida

pema@lsi.cefetmg.br

Computer Department, Federal Center for Technological Education of Minas Gerais

Av. Amazonas, 7675, Belo Horizonte, 30.510-000, Minas Gerais, Brazil

Abstract. *Patents are an important source of information for measuring the technological advancement of a specific domain of knowledge. The information contained in the patents is distributed in a large number of fields that can be accessed in the portals made available by the governmental authorities. Classification systems used by these offices group the documents into sections, classes, subclasses, groups and subgroups. An automatic system for identifying the set of keywords or key terms related to the main contents of patent groups would enable advances in the issues associated with the automatic generation of titles in patent subgroups. To implement this strategy, it is essential to identify the keywords that represent the patent since the keywords are not pre-defined by the patent authors. This work presents results of an empirical experiment, which compares the application of three algorithms for the extraction of keywords in patents, namely: X^2, TF.IDF and Wiki-TFIDF. The patent database used for this experiment consists of 100 patents from the United States*

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

Patent and Trademark Office (USPTO). The results of this research contribute to the organization of these databases and to the discovery of knowledge related to the information contained in the documents of patents made available digitally.

Keywords: *Classification Systems, Extraction of keywords, Patent*

INTRODUCTION

Patents are an important source of information for measuring the technological advances of a specific domain of knowledge. Authors such as Camus and Brancaleon (2003) highlighted the importance of the information contained in the analysis of patent documents, revealing the risks and opportunities of research and obtaining knowledge about the activities of the companies involved. Some studies (MARKELLOS et al., 2002; LEYDESDORFF, 2004) explore patent databases, showing how the production of scientific knowledge may be related to the economy. Lee, Yoon and Park (2009) stated that patents are complete sources of technical and commercial knowledge. The analysis of the technical and market attributes of patents has been considered as a useful tool for research and for managing development.

In order to extract knowledge from the information contained in 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. An alternative, object of study of this article, that would help the initial understanding of the documents content, would be the choice of keywords, similar to what is done in scientific articles. Due to the huge amount of patents already deposited in repositories around the world, creating these lists manually becomes impractical. The automatic acquisition of keywords, for example, would optimize the work of classifying patents by specialists of the offices managing the patenting process.

Keyword extraction has been explored in some previous work, on web pages, for determining relevant advertisements, refining database queries, summarizing, assisting article authors, and others. Siddiqi and Sharan (2015) presented some methods of extracting keywords that have been used by the scientific community. The methods were classified into four groups: statistical, supervised, semi-supervised and unsupervised.

Mihalcea and Csomai (2007) evaluated the application of unsupervised keyword extraction algorithms on wiki pages and the evaluation of word disambiguation algorithms. The ultimate goal was to build keyword links with Wikipedia's knowledge base. In the work presented by Reginaldo, Carvalho and Meireles (2016), an experiment was carried out to evaluate unsupervised algorithms for keyword extraction. In Vidal et al. (2012), three new methods for extracting keywords from web pages were presented using Wikipedia as a source of external information. The authors used various attributes of Wikipedia pages, such as article titles, co-occurrence of keywords and categories associated with each Wikipedia definition. The new methods were compared with three other extraction methods of keywords traditionally used in the literature.

In this article, we describe the steps of choosing and preparing the experimental databases and analyzing the results generated by the use of three algorithms for extracting keywords in patents. The following sections describe the patents, the classification systems used by the

offices and the algorithms used in the extraction of the keywords. Section 3 presents the methodology used and sections 4 and 5 present the results, analyses and final considerations.

1 THE PATENTS AND THE CLASSIFICATION SYSTEMS

A patent is a contract between the inventor and the government. In exchange for public and full disclosure of an invention, the government grants the inventor the right to exclude third parties for a limited period of time and to use or sell the invention (HUFKER; ALPERT, 1994). According to Alberts et al. (2011), patents are complex legal documents, which contain more details and descriptions than scientific articles. The language and formatting of the patent text are controlled by the laws and regulations of the country or by the patent authority in which the inventor applied for the patent.

To be eligible for patent protection, a product must meet three criteria: the invention must be new and different from previous inventions, it must be considered non-obvious and can not have been anticipated by any other invention (HUFKER; ALPERT, 1994). According to these authors, the United States patent office distinguishes three general patent classifications. The first class is that of utility and is related to a process, a machine, a manufacturing item or an improvement of one of these items. This type of patent is granted for a period of 17 years. The second class is related to a new and original design, and form is the main factor for a production item being considered patentable. Attributes such as mechanical utility and functionality are less relevant for this type of patent. This type of patent is issued for a period of three and a half, seven or fourteen years. The third class protects different inventions produced through biotechnology and not found in nature.

Table 1. Description of the fields available for query in the database of the United States Patent and Trademark Office (USPTO)

Field	Description
Abstract	A brief summary of the invention patent
Examiner	Name of the person responsible for examining the patent application
Claims	Text that presents the characteristics of what the applicant considers to be the invention and defines the scope of the claimed protection
Cooperative Patent Classification (CPC)	The classification code in which the patent was classified
Description	Detailed description
Title	Title of the patent
Classification	The original class and related classes in which the patent was classified are listed in the most recent classification files
Citing patents	List of other patents that contain this patent number in its text

Source: http://www.uspto.gov

The information contained in the patents is distributed in a large number of fields that can be accessed in the portals made available by the authorities or by means of search tools.

Fields are usually represented in an XML structure. Table 1 shows some of the main fields, and their descriptions, which can be used in the experimental stages of patent-related work.

There are important classification systems used by patent offices. These systems organize patents into categories according to their technical application, structural characteristics and use. The International Patent Classification (IPC), which contains the largest collection of patents available, is maintained by the World Intellectual Property Organization (WIPO) and divides all areas of technology into eight sections, subdivided into classes, subclasses, groups, and more than 60,000 subgroups. The United States Patent Classification (USPC) ranks patents in approximately 470 classes and 163,000 subclasses (ALBERTS et al., 2011). The European Patent Classification (ECLA) is maintained by the European Patent Office (EPO) and is a variation of the CPI. It has approximately 129,000 categories. The Cooperative Patent Classification (CPC) is a partnership between the USPTO and the EPO and was created to unify their respective classification systems, USPC and ECLA, including approximately 250,000 CPI-based categories. Table 2 shows the sections defined by the CPC.

Table 2. Sections defined by CPC

Sections	Descriptions
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G	Physics
H	Electricity
Y	General Tagging of New Technological Developments; General Tagging of Cross-Sectional Technologies Spanning over Several Sections of the IPC; Technical Subjects Covered by Former USPC Cross-Reference Art Collections and Digests

Source: http://www.uspto.gov

2 ALGORITHMS FOR EXTRACTION OF KEYWORDS

To extract keywords from patent documents, three algorithms were selected: TF.IDF, X^2 and Wiki-TFIDF. This work presents a comparison between these algorithms in the context of patent documents.

2.1 TF.IDF

The TF.IDF algorithm is an unsupervised method, which is based on the local (TF) and global (IDF) frequencies of the set or database items. The items will be treated as terms, since

the algorithms were applied to the texts. According to Baeza and Neto (1999), the frequency of the term (TF) is defined by Eq. (1), Eq. (2) and Eq. (3). Equation (1) calculates the importance of the term for the document by the frequency quotient of term i in document j by the frequency of the term that occurs most in the document.

$$f_{ij} = \frac{\text{frequency}_{i,j}}{max\ (\text{frequency}_{i,j})} \tag{1}$$

Equation (2) calculates the importance of the term for the collection of documents by the logarithm of the quotient of the total number of documents by the number of documents that have the term under analysis.

$$idf_i = log \frac{N}{n_i} \tag{2}$$

Finally, Eq. (3) calculates the weight of the term.

$$w_{ij} = f_{ij} \,.\, idf_i \tag{3}$$

2.2 X^2

The X^2 algorithm is, according to Manning and Schütze (1999), an independence test called the Chi-Square test or simply X^2. This test evaluates the independence between two variables and compares the observed values with the expected values. The test consists of evaluating how far these values are. In this work, as used by Mihalcea and Csomai (2007), the Chi-Square test is used to order the terms by their dependence on the patent, so that the greater the weight given by the test to a term, the greater its dependence to the document.

The X^2, in this work, is calculated on the variables "words" and "documents", both qualitative variables, which assume only two values as shown in Table 3.

Table 3. Values of the variables of the independence test

Variables	Words		Documents	
Values	m	All the other words	n	All the other documents

These variables are then placed in a Contingency Table, as shown in Table 4, so that the occurrences of a pair of values (words, documents) can be calculated.

The Chi-square test is defined by Manning and Schütze (1999) by Eq. 4 and Eq. 5. The observed values are the in bold values of Table 3, which are denoted by Oij where i is the line and j is the column of the Contingency Table. The expected values, denoted by Eij, are obtained by Eq. 4.

Table 4. Contingency Table of X^2

Words/ Documents	m	All the other words	
n	**Score (word m in document n)**	**Score (all the other words in document n)**	Sum (line)
All the other documents	**Score (word m in all other documents)**	**Score (all the other words in all other documents)**	Sum (line)
	Sum (coluna)	Sum (column)	Sum (Internal cells)

Adapted from MIHALCEA and CSOMAI (2007)

$$E_{ij} = \frac{\sum_a O_{aj} \cdot \sum_b O_{ib}}{\sum_{i,j} O_{ij}} \quad , \quad E_{ij} = \frac{\text{sum(column j)} \cdot \text{sum(row i)}}{\text{sum(total)}} \tag{4}$$

Finally, the value of the test is given by Eq. 5, where m is the word and n is the document. This calculation is redone for each word in each document.

$$\chi^2_{mn} = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{5}$$

2.3 Wiki-TFIDF

The Wiki-TFIDF algorithm was defined in Vidal et al. (2012). Wiki-TFIDF is a variation of TF.IDF. The Wiki-TFIDF method also uses the title database of the Wikipedia main namespace (22-Jun-2017), which lists the articles on the website and provides external information that refines the choice of keywords. In this method, the frequency of words in patent text is used as TF and the frequency of wiki links on wiki pages. Wikipedia is also used to create a controlled vocabulary that is used in the pre-processing of the patent base.

Wiki-TFIDF is defined by Eq. (6), where f_{ij} is the frequency of the term w in document d, N is the number of documents in Wikipedia, and n_i is the number of documents in Wikipedia where w occurs at least once. Thus, each word w receives a weight in each document d.

$$w_{ij} = \left(1 + log\left(f_{ij}\right)\right) \cdot log\left(\frac{N}{n_i}\right) \tag{6}$$

3 METHODOLOGY

In this section, the methodological phases of the work are presented, involving the creation of the used databases, the characteristics of the experiments performed and the metrics used to compare the results.

3.1 Databases

To conduct patent experiments, a database of 743 patents from a USPTO subgroup was created. Of these patents, 100 were selected from the G06K 7/1443 subgroup, called

"Recognition of data, presentation of data, record carriers; Handling record carriers ", which was called SG43_100P.

To run the Wiki-TFIDF algorithm, it is necessary to use the Wikipedia title database as a vocabulary. Thus, in order to construct a controlled and identical environment for the three algorithms, the list of all the titles of the main namespace of Wikipedia was used as the set of expressions of our vocabulary. This base was called Wiki-Voc.

The Wiki-Voc database also has, for each title, a counter of occurrences of that title in other documents in the form of a link. Thus, for each document D where one or more links to the title T appear, the counter of T is incremented.

3.2 Preprocessing

The chosen databases are composed of raw texts, requiring a preprocessing phase. Thus, a stage of preprocessing of texts was made to eliminate unnecessary components and to unite components of the same meaning. The process, presented in Figure 1 and Figure 2, aimed to preprocess all the bases used. In all bases, the processes of stop words removal and stemization were made. Vocabulary expressions are formed by n-grams. The preprocessed vocabulary is sent to the patent preprocessing stage, where it is used in the final process as a filter, ignoring any patent words that do not exist in the vocabulary. In a final stage, patent preprocessing removes vocabulary words that never occur on the basis of SG43_100P in order to eliminate irrelevant data.

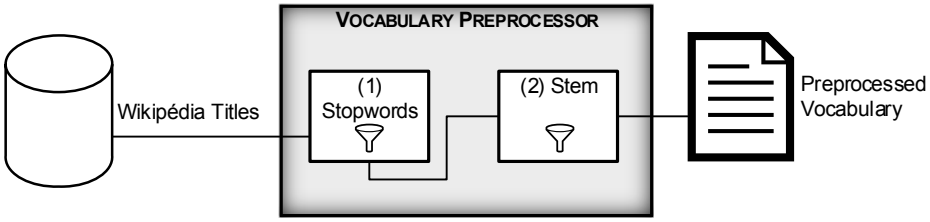

Figure 1. Vocabulary Preprocessing

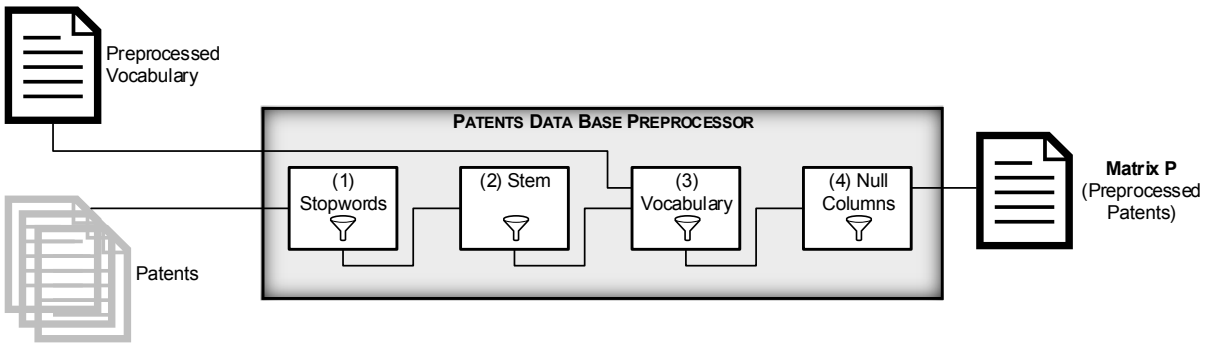

Figure 2. Patents Preprocessing

At the end of the preprocessing, the base SG43_100P is represented by an occurrence matrix P, shown in Figure 3. Each line i of the matrix represents a patent p_i where in each cell f represents the frequency at which a word w_j occurs at p_i.

$$P = \begin{bmatrix} f_{p_1,w_1} & \cdots & f_{p_1,w_m} \\ \vdots & \ddots & \vdots \\ f_{p_n,w_1} & \cdots & f_{p_n,w_m} \end{bmatrix}$$

Figure 3. Matrix P

3.3 Ground-truth

To validate the experiments, it was necessary to create a ground-truth, which for this work would be a base of predefined keywords. As no ancillary bases were found that provided keywords in patents, an alternative was proposed.

USPTO patents are classified by the CPC. In this system, there is a hierarchy of classifications, where each grouping receives a description. Similar to what was done in Reginaldo, Carvalho and Meireles (2016) and as shown in Figure 4, the descriptions from the root to the branch in which the subgroup G06K 7/1443 was located are concatenated, which contains the base SG43_100P. This set of concatenated descriptions was preprocessed as plain text and then generated a control base with 36 expressions, called GT_36.

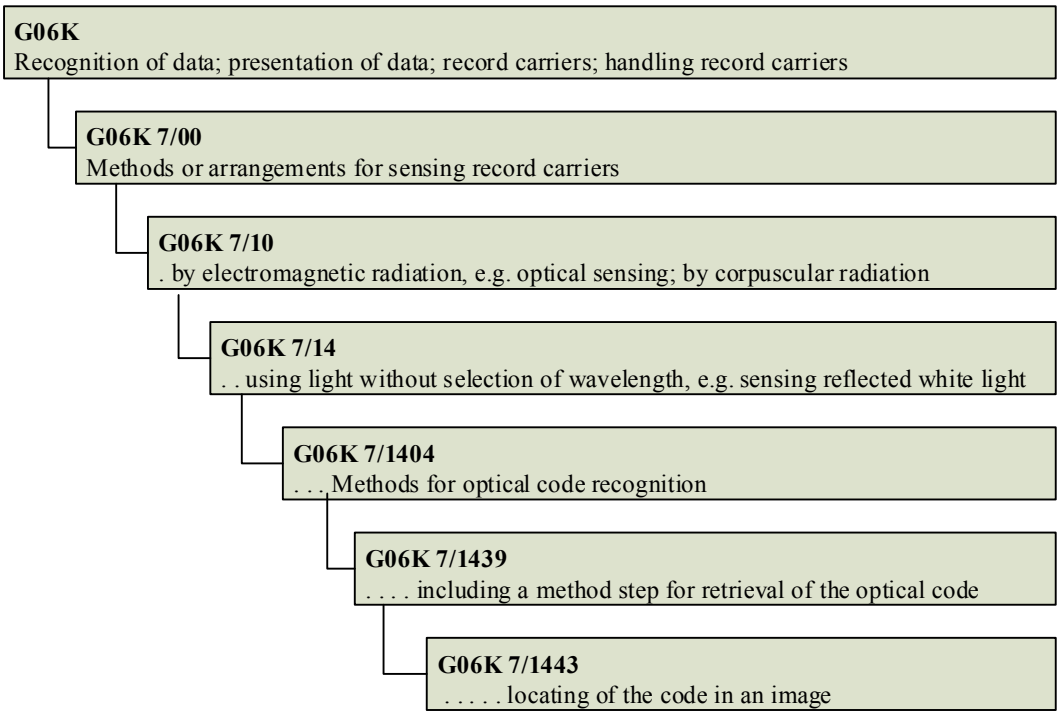

Figure 4. Descriptions of subclass G06K, which contains GT_36

Source: http://www.uspto.gov

GT_36 was considered as an appropriate set of keywords for the subgroup G06K 7/1443 because it is formed by the descriptions of its classifiers.

Preprocessing 1. The set of descriptions of subgroup G06K 7/1443 was preprocessed according to Figure 5, in three stages. In the first, the set was taken as a plain text and made

the stemization and generation of n-grams. In the second stage, the preprocessed Wiki-voc vocabulary is passed as a filter, removing any records that do not exist in the vocabulary, thus avoiding attempts that even exist in the vocabulary. In the third stage, a second filter was made, where it was verified which terms of the descriptions are present in the patents. At the end of these stages, the ground-truth GT_36 was formed with 36 keywords.

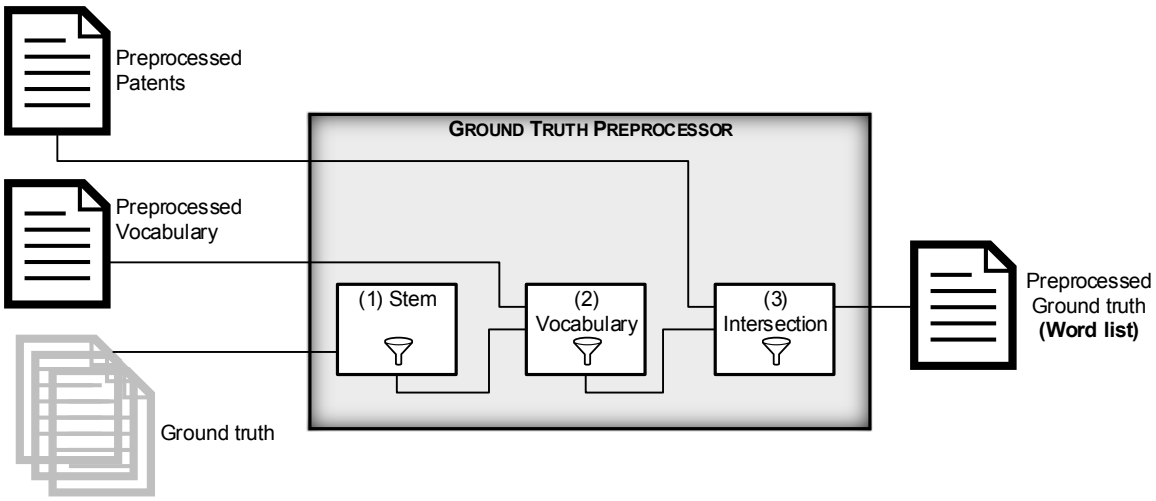

Figure 5. Ground-truth Preprocessing

Preprocessing 2. After some experiments in which there were low compatibility between the GT_36 and the results of the algorithms used in this work, it was noticed that, although they exist in the patents, some items of GT_36 presented low frequency. Thus, a new ground-truth was generated from the previous one, identifying the words of GT_36 that were repeated more frequently (more than 8 times) in each patent of base SG43_100P. Then, a ground-truth with 15 keywords, denoted by GT_15, was created.

Ground Truth based in Titles. As a second alternative of comparison, another ground truth, named GT_Titles, was also created, which is formed by the titles of the 100 patents of the base SG43_100P.

3.4 Experiments

From the preprocessed databases, that is, from the P matrix representative of the base, as previously described, three experiments were undertaken for each of the algorithms: X^2, TF.IDF, Wiki-TFIDF, shown in Figure 6.

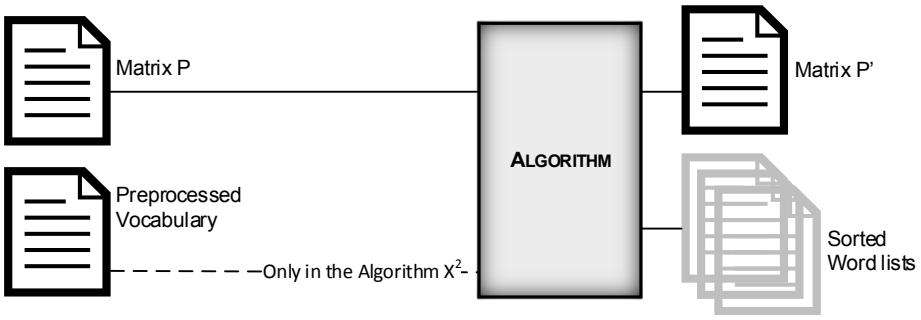

Figure 6. Experiments with patents base

Each algorithm receives the matrix P and returns another matrix, P', of the same dimensions still representing the patents. Matrix P' has, in each cell, weights h for the words w in each patent p. Thus, each word w of p has a weight h, which will be used to rank them in each patent. As Figure 6 shows, each row of the matrix is transformed into a list of words w ordered by the weights of h, which we call L_{wTotal}. All experiments followed this initial processing. The following stages of the experiments are differentiated and processed as follows:

Experiment 1 – Ground truth GT_15. From P', a list of L_{w15} words ordered by their weight was generated for each patent, in which each list is limited to only 15 words with greater weight, aiming to reconcile the cardinality of the recovered data with the ground-truth cardinality GT_15.

Experiment 2 - Ground truth GT_Titles. Similarly to experiment 1, word lists L_{wx} were generated in which x represents the number of words in the GT_Titles for the patent whose list of words is in focus. This aims to reconcile the cardinality of the data retrieved with the cardinality of the ground truth GT_Titles

Experiment 3 – Comparison between the algorithms. In this experiment, the L_{wTotal} lists are maintained and they are then compared through the Rand-Index, in order to identify similarities among the algorithms. To compare the algorithms two by two, we compared each L_{wTotal} word list of a patent p of an algorithm x with the respective list L_{wTotal} of the same patent p of an algorithm y.

3.5 Validation Metrics

To validate the results, the Rand-Index, Precision and Revocation metrics were used.

Rand-Index. Rand-Index is used to compare two clusters originating from the same database. In this metric, the original set is the base SG43_100P and clusters are the lists of words that each algorithm returns from P'. Thus, an assessment can be made of how similar the algorithms are in selecting keywords. The Rand-Index is given in percentage, and a value of 100% means total similarity between the results of the algorithms.

Precision and Revocation. Precision is calculated by dividing the number of words correctly identified as keywords by the number of words proposed by the algorithm. Revocation is defined as the division of the number of keywords correctly identified by the number of keywords defined by the ground-truth for the document. To consider the calculation of the metrics for the sub-group of patents in the base SG43_100P, mean, standard deviation, minimum and maximum values of precision and recall of patents in the base SG43_100P

were calculated. To calculate these metrics, each L_{P15} list is taken as the set of recovered data, and the ground truth as the true set to be retrieved in each experiment.

A 100% recall means that all ground truth has been correctly recovered, and 100% accuracy means that of all the recovered data, all belong to ground-truth.

4 DISCUSSION OF RESULTS

The experiments were performed and the results evaluated using the metrics presented in section 3. In all the evaluations, mean, standard deviation, and minimum and maximum values are presented.

4.1 Experiment 1 - Ground-truth GT_15

The word lists L_{w15} were compared with the ground truth GT_15. According to Figure 7, it can be seen that X^2 achieved 13% recall and accuracy at most, which means having up to 2 of 15 words hit. TF.IDF achieved 6% recall and accuracy at most, which means that it has scored up to 1 in 15 words. Wiki-TFIDF failed to achieve success in any of the 100 patents on this ground-truth.

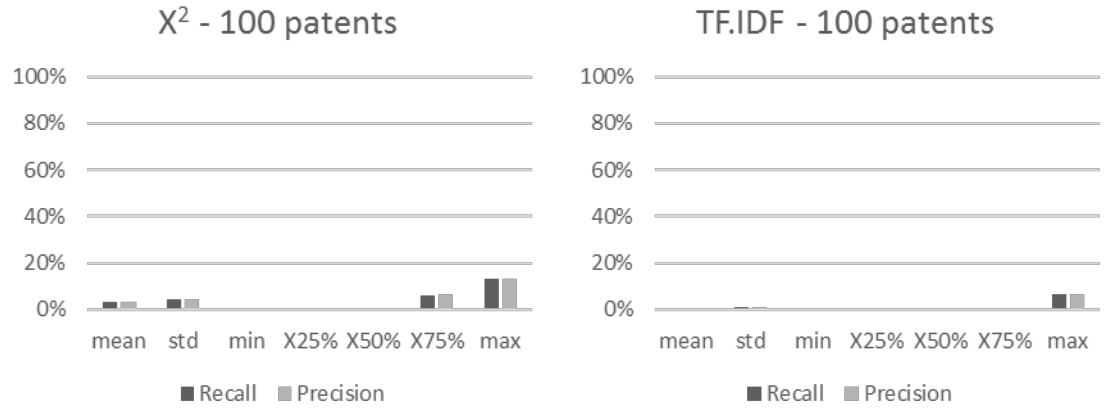

Figure 7. Results of experiment 1, algorithms X^2 and TF.IDF at the base SG43_100P compared to GT_15

The algorithms presented little significant results in relation to the ground truth GT_15, which creates the hypothesis that the ground-truth created is not an appropriate option for this database.

4.2 Experiment 2 – Ground truth GT_Titles

The L_{wx} word lists were compared to the ground-truth GT_Titles. As shown by Figure 8, it is observed that X^2 achieved 44% recall and 67% accuracy at most, and 3% on average for both metrics. The TF.IDF, however, achieved 9% recall and 11% accuracy on average and, respectively, 50% and 100% maximum. Wiki-TFIDF achieved 7% recall and 25% accuracy at most.

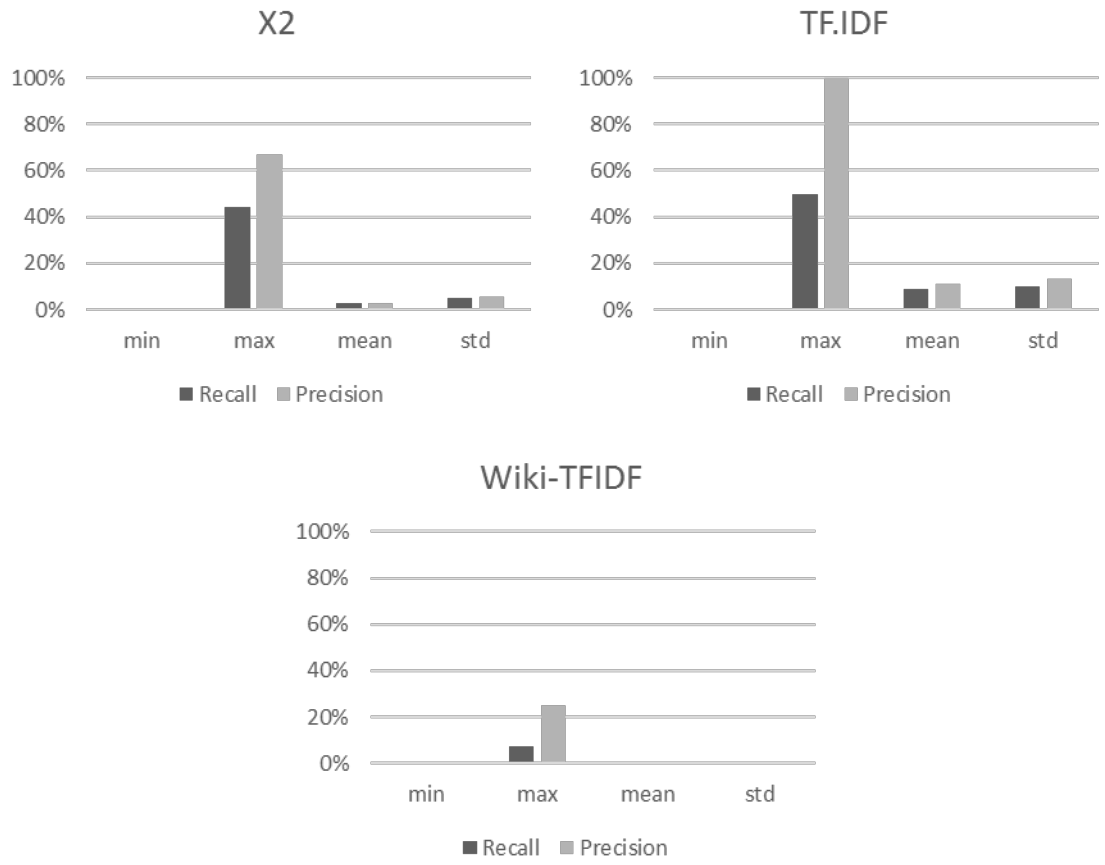

Figure 8. Results of experiment 2, algorithms X^2, TF.IDF and Wiki-TFIDF at the base SG43_100P compared to GT_Titles

The algorithms presented slightly better results when the ground truth GT_Titles was used in relation to the ground truth G_15.

4.3 Experiment 3 – Comparison between the algorithms

The algorithms X^2, TF.IDF and Wiki-TFIDF were compared, two by two, using the Rand-Index metric, as shown in Figure 9.

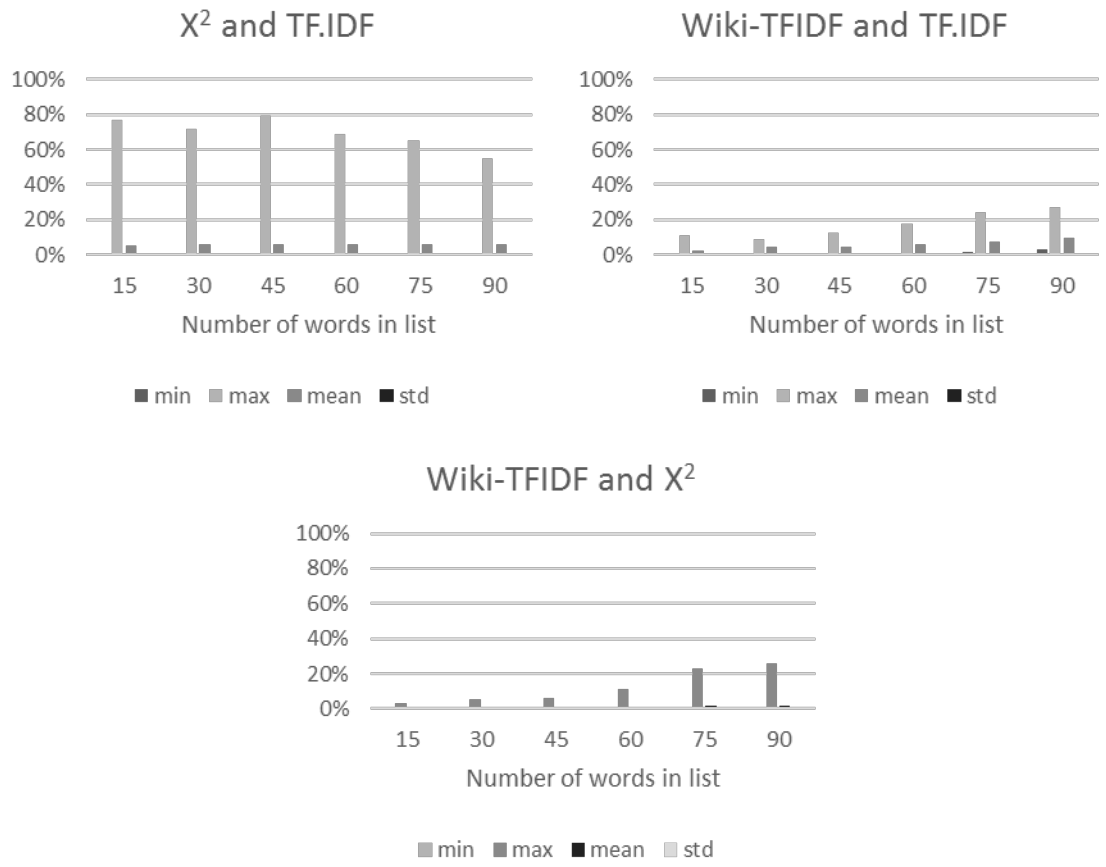

Figure 9. Results of experiment 3, comparison between algorithms X^2, TF.IDF and Wiki-TFIDF at the base SG43_100P

Analyzing the Rand-Index, we conclude that the TF.IDF and Wiki-TFIDF algorithms have greater similarity. However, the methods TF.IDF and X^2 have a small similarity against an almost total dissimilarity between Wiki-TFIDF and X^2.

5 CONCLUSION

In this work, a comparison of three algorithms, TF.IDF, X^2 and Wiki-TFIDF, for the extraction of keywords in patent documents was proposed. To validate the results found, lists of words that could constitute the ground truth of the experiments were formulated. Some alternatives have been presented through descriptions of the patent classifications used by USPTO and patent titles. In these experiments, the algorithms TF.IDF and X^2 found a larger number of words compared to the proposed ground truth. When comparing the results of the algorithms, using the Rand-Index, these two algorithms also found greater similarity.

Whereas even the TF.IDF algorithm, traditionally used by the scientific community, obtained insignificant results, it is assumed that the ground truth chosen is not suitable for the task, requiring, for example, the use of a manual ground truth, prepared by specialists. This emphasizes the importance of prior construction of a data validation base.

The automatic generation of a set of keywords would enable advances in research associated with the generation of titles of new groups generated in patent classification systems or in the reclassification process required when a large number of patents in a given group is reached. For the continuity of this work, it is suggested, first of all, the manual creation of a ground truth with the help of specialists in the area of knowledge approached by the patents database.

ACKNOWLEDGEMENTS

We thank the Pontifical Catholic University of Minas Gerais – PUC Minas and the National Council for Scientific and Technological Development - CNPq for funding this research.

REFERENCES

Alberts, D., Yang, C. B., Fobare-DePonio, D., Koubek, K., Robins, S., Rodgers, M., Simmons, E., & DeMarco, D., 2011. Current Challenges in Patent Information Retrieval. In Lupu, M., Tait, J., Trippe, A. J., & K. Mayer, eds, *Introduction to Patent Searching*. pp. 3-44. Springer.

Baeza-Yates, R., & Neto, B. R., 1999. *Modern information retrieval.* vol. 463. New York: ACM press.

Hufker, T., & Alpert, F., 1994. Patents: A managerial perspective. *Journal of Product & Brand Management*, vol. 3, n. 4, pp. 44-54.

Leydesdorff, L., 2004. The university–industry knowledge relationship: Analyzing patents and the science base of technologies. *Journal of the Association for Information Science and Technology*, vol. 55, n. 11, pp. 991-1001.

Lee, S., Yoon, B., & Park, Y., 2009. An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, vol. 29, n. 6, pp. 481-497.

Manning, C. D., & Schütze, H., 1999. *Foundations of statistical natural language processing.* Cambridge: MIT press.

Markellos, K., Perdikuri, K., Markellou, P., Sirmakessis, S., Mayritsakis, G., & Tsakalidis, A., 2002. Knowledge discovery in patent databases. In the *Eleventh international conference on Information and knowledge management (CIKM'02),* pp. 672-674.

Mihalcea, R., & Csomai, A., 2007. Wikify!: linking documents to encyclopedic knowledge. In *Sixteenth ACM conference on Conference on information and knowledge management*. pp. 233-242.

Reginaldo, T. V., Carvalho, J. R., Meireles, M. R. G., 2016. Automatic identification of keywords to generate links in patent documents. *V Simpósio Mineiro de Sistemas de Informação.* pp. 185-196.

Siddiqi, S., & Sharan, A., 2015. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, vol. 109, n. 2. pp. 18-23.

United States Patent and Trademark Office. http://www.uspto.gov/.

Vidal, M., Menezes, G. V., Berlt, K., de Moura, E. S., Okada, K., Ziviani, N., ... & Cristo, M., 2012. Selecting keywords to represent web pages using wikipedia information. In the *18th Brazilian symposium on Multimedia and the web*. pp. 375-382.