Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 78 (2016) 439 – 446

Procedia Computer Science

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

Keyword based search and its limitations in the Patent document to secure the idea from its infringement

Ranjeet Kumar[a]* , R.C.Tripathi[b], Vrijendra Singh[c]

[abc]*Indian Institute of Information Technology, Allahabad, Deoghat Jhalwa, Allahabad, U.P*

Abstract

Intellectual Properties (IP's) are attracting progressively growing popularity for corporate houses and the academia in the current years. Patent system is one of them which generate high economical values of the IP rights. This in turn calls for the increased work responsibility of patent prior art search to generate effective patent search reports for the innovator (s). In the field of patent innovations, prior knowledge of innovative steps of the technologies developed so far must be known to innovator (s). In the present research work, technology/ patent search based on keywords has been investigated to arrive at the usefulness of the methodology particularly for the case of patent documents. The present paper helps to figure out the limitations and the scope of the methodology for patent prior art search based on extent of the keywords.

Peer-review under responsibility of organizing committee of the ICISP2015

Keywords: Patent Novelty Search; Text Search; Keywords Based; Prior Art Search; Patent innovation Search

1. Introduction

Patents provide ample resource of the knowledge of technology in the specific area of research. They contain both legal and technological boundaries in their unique piece of document in a typical area of research. The patent document may provide privacy of the research done from infringement of idea/research as per non-violation legal boundaries it contains. The term privacy is used in terms of technology infringement or legal protection granted by territorial authority from infringement of the patented idea for the certain period of time. It may be called the

* Corresponding author. Tel.: 0532-292-2161; fax: +0-000-000-0000 .
*E-mail address:*ranjeet@iiita.ac.in

Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.086

patented technology is a private document for inventor. For non-infringement case of patent document patent researches performs different type of patent searching at different stages of patenting process. Patent search is a general term that covers different types of search processes such as for technology survey, prior art search, freedom to operate, validity and patent portfolio search etc. These search processes differ in terms of the information need of the searcher, the corpora and the output of the search. Notice, however, that the precise names and definitions of these search processes vary between those who deal with patents, like for example, information specialists, private patent searchers, patent examiners, and patent lawyers (Lupu and Hanbury 2013)[1].

Many researches are ongoing in the area of patent information retrieval for different purposes of research. Patent research analysis [R.Kumar, et. al 2011][2] and their importance in many ways in the global economy has been prime area of research in current scenario. Patent prior art search is one of the known research problems in this domain. Patent search is an active sub-domain of the research field known as information retrieval [IR; Tait 2008][3]. The prior art search is the common task before filing a patent for grant. The inventor(s) as well as the patent examiner perform the prior art search for assuring the originality of the invention by using information retrieval techniques. These searches are performed using the keywords or phrases or some set of words extracted from the draft patent application, which are used typically as the query [Mahdabi et al. 2012; Piroi et al. 2011; Xue and Croft 2009] [4, 5, 6]. Keyword based query search is the simple query search which searches general documents in which it must handle the query expansion problems for the patent prior art search. However, this problem is more critical in patent search than in the search of general research documents. In the present paper, the search process based on keywords for a new patent document has been investigated to figure out the scope of the methodology and its limitations.

2. Literature Survey

A patent document search represents long technical document as a query for the search performance. Early systems mimicked the approach taken by professional patent examiners, who (at the time) valued high frequency words as query terms (Itoh et al. 2003; Iwayama et al. 2003)[7, 8]. Generating the appropriate query for keyword based extraction methods has been most utilized to make a retrieval process for the patent prior art search as the basic concept. The objective herein is to find more relevant literature based on these extracted keywords. In their research [Magdy et al. 2009][9] focused on the query length and stated that patent applications demand queries comprising hundreds of words as opposed to ad-hoc web searches where queries are usually rather short. [Xue and Croft 2009] experimented this with the data taken from United States Patent and Trademark Office (USPTO). They examined using query terms taken from different fields of a patent document. In the experiment they found the best performance was obtained when using high frequency terms extracted from the raw text of the “Description” field. In their research [Mase et al. 2005][10] proposed a two-stage patent retrieval method, in which the first stage uses the similar approach to expand the query, and the second stage focuses on claim parsing to re-weigh the query words and then identify the top 1000 patents. In the method of keyword extraction an expansion focuses on the Key phrase Extraction from Scientific Articles, [Lopez and Romary, 2010b][11]. They have proposed to use up to 5-grams for the phrases extraction from textual content of the document. In the researches in this area there are various state-of-the-art patent IR systems [e.g., Becks et al. 2011; Lopez and Romary 2010; Magdy and Jones 2010; Mahdabi et al. 2011][12, 13, 14]. All of these systems use single query representations of the patent application. In contrast described below is our approach for prior-art search that uses multiple query representations.

3. Search Methodology

A text search is performed by using one or more search keywords to query bibliographic data, indexed data, and sometimes abstract and even full text data in an electronic database. Text searching is often aided by special operators. Widely known operators include the Boolean AND, OR, and NOT operators, but may also include proximity operators that specify the order between two words and the maximum allowed distance that should exist between them. Allowed operators may vary depending on the search engine that is selected. One benefit to a text search is that it can find “outlying” documents that have been improperly classified. A global text searching strategy should be used independent of classification limitations wherever possible to ensure that the misclassified documents get a higher chance of being examined during the search. Identifying Keywords is one of the biggest obstacles for an effective text search. The need is to identify all potential keyword combinations that could describe the search subject matter. When identifying keywords related to the search subject, it is vital to consider the function

of the invention/product as well as its component parts.

3.1. Keyword based search

The terms used in the patents are quite unlike the other documents like newspaper or scientific articles wherein many vague or general terms are often used in order to avoid making/ narrowing down the scope of the observations. Combination of general terms may contain a special meaning. To identify these combinations are important. At the same time, patent documents include many acronyms and new terminology. For the patent searching on the basis of keywords matching, it is important to make the text free from stop-words and other frequently used texts. One needs here to create a data-bin where most important keywords stored from the document and then apply the words similarity algorithm for the patent searching.
Use of keywords for initial search is the most frequently employed approach, in comparison to the use of "classification code" to initiate search. The latter is less often adopted by patent engineers. Use of synonyms for search vocabulary is the second most often employed method in patent search. It is followed by "use of related words for search vocabulary," and then by "use of broader terms (hyponyms). The last choice is "use of narrower terms (hyponyms) for search vocabulary. This implies synonyms come first as the search vocabulary develops.

3.2. Keyword association based search

In the methodology of the keyword association based search, the keywords have to be generated from the document and preprocessed. At First, one eliminates the stop words from the text. Second step is to execute stemming process to route the keywords arrived at relevant search results. After that, one identifies all the keywords from the text. Now, statistical analysis, frequency analysis in particular, is conducted first to set individual weights of words and weights of respective relations of each other. Then, word stems with high frequency are determined as keywords. Finally, keyword vector is constructed.
After getting keyword vector, the incidence matrix is constructed as a prerequisite for generating the search network. To construct the incidence matrix, the relationship between patents should be quantified in terms of either distance or similarity. Among various association indices, the common Euclidian distance index is used in this research [Johnson and wichem, 1988; R Kumar et. al, 2013][17, 18]. If keyword vectors of *ni* and *nj* are defined as (ni_1, ni_2, …, *nik*) and (nj_1,nj_2, …, *njk*), respectively, the association value between two vectors (nodes) is computed as follows:

$$Association\ Value = \sqrt{\frac{(n_{i1}-n_{j1})^2+(n_{i2}-n_{j2})^2+\cdots+(n_{ik}-n_{jk})^2}{k}} \qquad (3.1)$$

Although the association values assume real numbers from 0.0 to 1.0, the incidence matrix contains binary values where entry *Iij* equals 1 if node *i* is strongly connected with node *j* but equals 0 if node *i* is not or is loosely connected with node *j*.

3.3. String matching search

For the query generated after the preprocessing of the abstract and important words and their associated terms, these are stored in a data-bin for the matching of the other patent documents for the patent searching process. Rabin-Karp algorithm has been used for the purpose. The algorithm is an improvement of the brute-force approach to pattern matching. This algorithm is a probabilistic algorithm that adapts hashing techniques for string searching. It uses extra memory to advantage by treating each possible m-character section of the text string (where m is the pattern length) as a key in a standard hash table, computing the hash function of it, and checking whether it equals the hash function of the pattern.
Here the hash function is defined as follows

$$h(k) = k \bmod q, \qquad (3.2)$$

Where q is a large prime number
A large value of q makes it unlikely that a collision will occur. It translates the m-character into numbers by packing them together in a computer word, which then treats as the integer k in the function above. This corresponds to writing the characters as numbers in a radix d number system. Where d is the number of possible characters. The

number k corresponding to the m-character section text[i]……text [i+m-1] is

$$k = text[i] * d^{m-1} + \text{text[i + 1]} * d^{m-2} + \cdots + \text{text[i + m} - 1] \quad (3.3)$$

Shifting one position to the right in the text string simply corresponds to replacing k by

$$(k - text[i] * d^{m-1}) * \text{d} + \text{text[i + m]} \quad (3.4)$$

4. The Results Analysis

The results of the keyword search have been analyzed on the databases of USPTO[15] (United State Patent and Trademark office) and EPO [16](European Patent office) respectively. The keywords have been generated using with the patent abstract of 'Information System' e.g. given below- the black text is most appropriate keywords of the abstract given. The search process is results performed on the said databases.

US patent No- 9147297

Abstract- Methods and systems for a complete vehicle ecosystem are provided. Specifically, systems that when taken alone, or together, provide an individual or group of individuals with an intuitive and comfortable vehicular environment. The present disclosure includes a system that provides various outputs based on a user profile and determined context. An output provided by the present disclosure can change a configuration of a vehicle, device, building, and/or a system associated with the user profile. The configurations can include comfort and interface settings that can be adjusted based on the user profile information. Further, the user profiles can track health data related to the user and make adjustments to the configuration to assist the health of the user.

Total word count of the abstract- 123

Maximum relevant keywords are- 54

The keyword search of 'information system' in the USPTO database generates 57154 results found as shown in Figure 1 below.

USPTO PATENT FULL-TEXT AND IMAGE DATABASE

Home | Quick | Advanced | Pat Num | Help

Next List | Bottom | View Cart

Searching US Patent Collection...

Results of Search in US Patent Collection db for:
"information system": 57154 patents.
*Hits **1** through **50** out of **57154***

Next 50 Hits

Jump To

Refine Search "information system"

	PAT. NO.	Title
1	9,241,327	LTE enhancements for small packet transmissions
2	9,241,302	Methods and apparatus for radio access technology search
3	9,241,287	Narrow bandwidth operation in LTE
4	9,241,274	Method for avoiding in-device coexistence interference in wireless communication terminal, terminal and eNodeB
5	9,241,259	Method and apparatus for managing the transfer of sensitive information to mobile devices
6	9,241,252	Identifying an entity associated with wireless network access point
7	9,241,187	Method and system for customizing television content
8	9,241,156	Method for estimating the type of the group of picture structure of a plurality of video frames in a video stream
9	9,241,144	Panorama picture scrolling
10	9,241,109	Image capturing apparatus, control method, and recording medium for moving image generation

Figure 1 shows the results produced in USPTO of keyword 'information system'

The test case of keyword based search in the European patent database found these typical results. When applied the 'smart search' option in the database search, total100, 000 results found based on the search in the title, abstract and bibliographic information. Figure 2 shows the results found in the European patent database

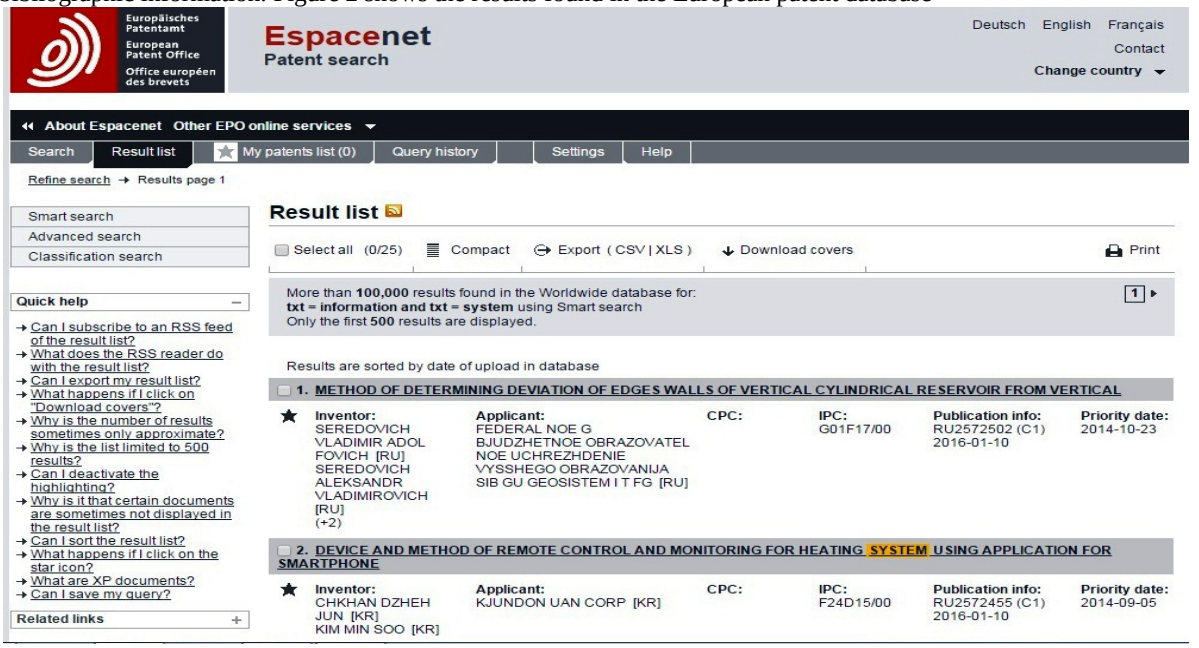

Figure 2 shows the results found in European patent database for information system

Test results of the combination of the keywords e.g. 'complete vehicle ecosystem' in the general search option of the USPTO database. The figure 3 shows the results found. Total 33 results found for the combination of the keywords.

Figure 3 shows the combination of keywords search in general search option in the USPTO database

Total 19 search results were found and the first result is the main patent document when the combination of the keywords in specific **abstract** search option is made in the USPTO. Figure 4 given below shows the results found.

Figure 4 shows the results found in the specific abstract search option in the USPTO database

The test case of combination of keywords used in the abstract in the patent 'WO2015141965 (A2)- System for providing Advertisement Exposure Information' in the patent databases of European patent database found these typical results. When applied the 'advanced search' (with title and abstract combination) option in the database search, total 35 results were found which based on the search in title, and abstract information, first results of the relevant patent document. Figure 5 shows the results found in the EPO for the said combination of keywords.

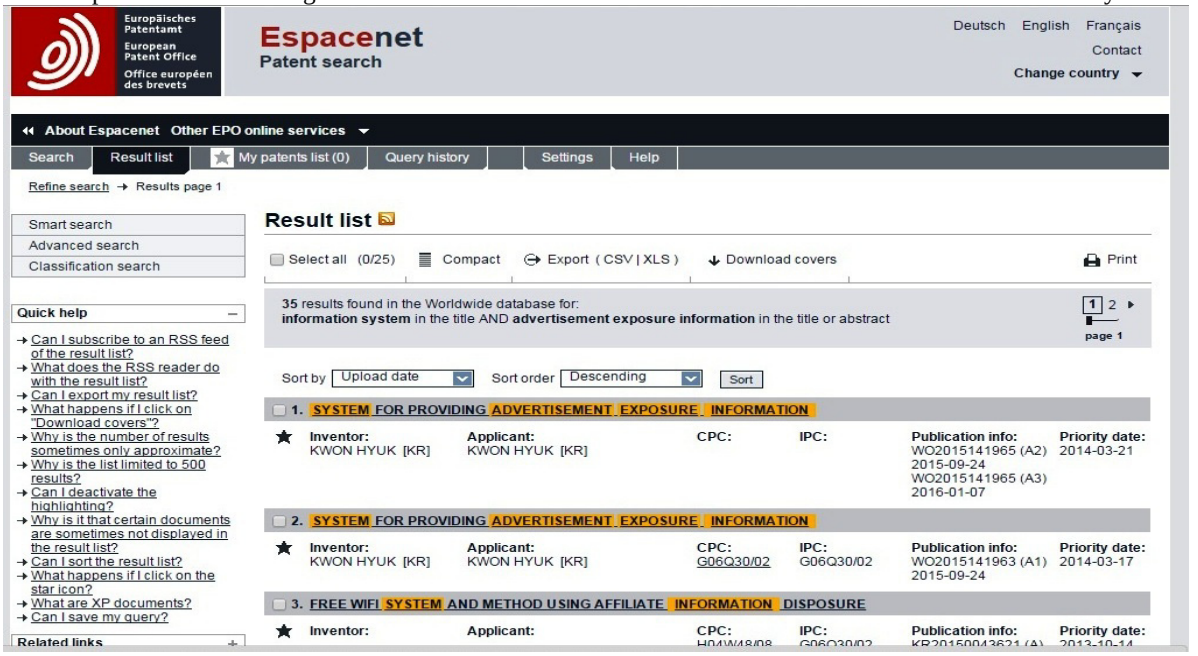

Figure 5 shows the results found of the combination of keywords in the EPO database

5. The Discussion

In the approach discussed in the present paper, there are some limitations which close boundaries of our methodology. In the USPTO web database, full-text keyword searching is allowed only for patents granted finally after 1975. In the case of patent applications, published after March 14, 2001 for the cases still to reach the stage of final grant facilities are available for keyword searching. In the US-PTO databases, logical keyword searches are possible only for a controlled vocabulary.
As such the keywords search produces huge set of results irrespective of whether these are concerned to the user or not. In the results obtained in the test cases of 'information system' more than 5000 search results were obtained both in the USPTO and EPO databases. The common phrase "Garbage in Garbage Out" (GIGO) applies to all patent databases. In that situation, the combinations of keywords only may produce relevant search results. In the test results of combination of keywords, the search results obtained relevant and reduced number of cases in both USPTO and EPO databases, because words form the weak link in patent searches. The only way to overcome this problem is to perform a classification search using one of the standard national/international systems e.g. USPC, CPC, IPC or F-Term and F-Index. This limitation needs to be observed by the independent inventor. The inventor must understand that their "research" is merely a preliminary search only.

6. The Conclusion

In the patent search problem, keywords based search is widely used. Many uses of this application have been done for the text searching and retrieval of data from different resources whether it is from the internet or from database repositories. The present paper investigated about the keyword based search and their efficacy in the vast patent databases e.g. USPTO and EPO. The test results show that the keyword based search results give enormously large seat of data. To resolve the problem, when combination of keywords has been used, then search results are more focused, reduced in number and accurate. This pattern was obtained in both USPTO and EPO databases. Thus to overcome of the limitations of the keywords based search, associations rule of the keywords and classification based search has been introduced. The string based search is the most accurate in terms of patent searching problems.

References

1. Lupu, M., & Hanbury, A. Patent retrieval. Foundations and Trends in Information Retrieval, 7(1), pp 1–97, 2013.
2. R, Kumar, R.C.Tripathi, M.D.Tiwari, A Case Study of Impact of Patenting in the Current Developing Economies in ASIA", Springer Scientometrics 88: 575-587, 2011
3. Tait, J. (Ed.). Proceedings of the 1st ACM workshop on patent information retrieval, PaIR 2008, Napa Valley, California, USA, October 30, 2008. ACM.
4. Mahdabi, P., Andersson, L., Keikha, M., & Crestani, F. Automatic refinement of patent queries using concept importance predictors. In Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12, New York, NY, USA: ACM, pp. 505–514, 2012.
5. Piroi, F., Lupu, M., Hanbury, A., & Zenz, V. CLEF-IP 2011: Retrieval in the intellectual property domain. In CLEF (Notebook Papers/Labs/Workshop), 2011.
6. Xue, X., & Croft, W. B. Transforming patents into prior-art queries. In Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR '09. New York, NY, USA: ACM, pp. 808–809, 2009.
7. Itoh, H., Mano, H., & Ogawa, Y. Term distillation in patent retrieval. In Proceedings of the ACL- 2003 workshop on patent corpus processing—volume 20, PATENT '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 41–45, 2003.
8. Iwayama, M., Fujii, A., Kando, N., & Takano, A. Overview of patent retrieval task at NTCIR-3. In Proceedings of the ACL-2003 workshop on patent corpus processing—volume 20, PATENT '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 24–32, 2003.
9. Magdy, W., & Jones, G. J. F. Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. In CLEF (Notebook Papers/LABs/Workshops), 2010.
10. Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., Oshio, T. Proposal of two-stage patent retrieval method considering the claim structure. *ACM Transactions on Asian Language Information Processing (TALIP), Volume 4 Issue 2, 2005.*

11. Lopez, Patrice and Laurent Romary, HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. In *SemEval 2010 Workshop*. Uppsala, Suede, 2010b.
12. Becks, D., Eibl, M., Ju¨rgens, J., Ku¨rsten, J., Wilhelm, T., & Womser-Hacker, C. Does patent IR profit from linguistics or maximum query length? In CLEF (Notebook Papers/LABs/Workshops), 2011.
13. Lopez, P., & Romary, L. Experiments with citation mining and key-term extraction for prior art search. In CLEF (Notebook Papers/LABs/Workshops), 2010.
14. Mahdabi, P., Andersson, L., Hanbury, A., & Crestani, F. Report on the CLEF-IP 2011 experiments: Exploring patent summarization. In CLEF (Notebook Papers/Labs/Workshop), 2011.
15. US Patent and Trademark Office (USPTO) http://patft.uspto.gov/netahtml/PTO/search-bool.html
16. European Patent Office (EPO) http://worldwide.espacenet.com/
17. R. Johnson, D. Wichern, Applied multivariate statistical analysis, Prentice Hall, Englewood Cliffs, NJ (1988)
18. Ranjeet Kumar, R.C.Tripathi, An Analysis of Automated Detection Techniques of Textual Similarity in Research Documents, International Journal of Advanced Science and Technology, Vol. 56, 99-110, 2013