# Measuring patent similarity by comparing inventions functional trees

**Gaetano Cascini**[1], **Manuel Zini**[2]

[1] University of Florence, Italy, gaetano.cascini@unifi.it
[2] drWolf srl, Italy, mlzini@drwolf.it

**Abstract:** The estimation of the conceptual distance between patents is a critical issue for Computer-Aided patent portfolio analysis systems, an emerging class of computer tools for supporting R&D analyses and decisions, patent infringement risk evaluation, technology forecasting. The aim of the present work is the introduction of an original algorithm for patent comparison: since typical text analyses are biased by the writer's style, the inventions similarity is here estimated by comparing the components and their hierarchical and functional interactions automatically extracted by means of a custom software tool. The whole procedure is clarified with an exemplary application in the field of electric current circuit breakers.

**Keywords:** Patent mining, document similarity, plagiarism

## 1. Introduction

Computer-aided patent portfolio analysis is an emerging topic in the scientific community and attracts interests from several disciplines, since it deals with economical, technical, management, life science issues [1-4].

Indeed, computers have been used for patent searches and analyses since the '90s, but most of the applications were limited to statistical computations by means of bibliometric methods. Indeed, these techniques are still adopted as a relevant source of information [5]. This is mainly due to a heritage of traditional practices when statistical techniques were adopted to examine the effect of technology development in economic, national and international contexts or to plan a corporate technology activity at a corporate level [2].

The introduction of text-mining algorithms has created new opportunities for identifying complex relationship among patent documents. Besides, up to now, the researchers in this field have dedicated major attention to Information Extraction purposes in order to capture relevant information from patents, while still limited

studies exist about patent comparison and trend extraction, except applications of general purpose clustering algorithms [6].

Nowadays, computer-based systems for patent analysis are assuming more and more specialistic roles and will cover soon a wider range of application areas like:

- generation of new research directions for biomedical studies [1];
- comparison of the morphology portfolio of different technologies [2];
- evaluation of the R&D landscape and business opportunities [3];
- evaluation of the risk of patent infringement [4].

All the above mentioned activities require the estimation of the conceptual distance between patents, but all the approaches proposed so far are based on keywords comparisons (e.g. co-occurrences of terms and/or multi-words), while the nature of patent contents is poorly taken into account. An even more critical issue is the dependence of these techniques to the language style of the writer; as a result, very often it happens that patents of the same inventor or company are clustered together despite their different contents, while conceptually close inventions are considered distant from each other just because they adopt a different terminology.

In the present paper the authors, also thanks to previous experiences in the field of plagiarism detection, propose a novel technique for assessing patent similarity as a means for avoiding the impact of the language style on patent comparison.

In the next chapter we report a brief survey of plagiarism detection techniques; then the third section describes the proposed algorithm also resuming some previous works relevant for the present application. Then an exemplary application of the proposed similarity metric is shown, by comparing the results of the automatic analyses performed by means of a prototype software to the results obtained by humans in the field of electrical circuit interruption devices. Finally, in chapter five, the discussion is focused on the capabilities and the sore points of the proposed technique.

## 2. State of the art techniques for plagiarism detection

Plagiarism is a growing problem and has recently received a lot of attention. The increase in availability of material in digital form has made plagiarism much easier.

However, the use of digital media also means that there are greater opportunities to trace plagiarism by means of dedicated software tools. Automated plagiarism detection as a subject has not yet achieved the same degree of scientific maturity as other subjects in the Text-Mining field, but a growing number of publications [7], websites and recently available products on this matter [8, 9] indicates that both the scientific and the industrial communities have started to recognize and acknowledge the existence of a recent problem which is yet awaiting its systematic solution [10].

There are several approaches to automatically identify plagiarism in different types of documents. The SCAM tool developed by Shivakumar [11] is based on building unions of word sets and counting domain-specific keywords in them. Plagiarism is then revealed via unexpected or otherwise suspicious occurrences of such keywords.

In some works, plagiarism detection has been regarded as a special case of duplicate document detection, which is both a necessary and difficult task in the management of large scale and very large scale databases (possibly multi-media databases). A variety of data mining methods and text-based techniques for such purposes have been proposed and investigated [12].

Comparing whole document checksums is simple and suffices for reliably detecting exact copies; however, detecting partial copies is subtler; in some works, for example in [13], an approach based on multiple fingerprints evaluation is used to detect partial copies. These techniques mostly rely on the use of k-grams, i.e. contiguous sub-strings of characters with length k. The process requires to divide a document into k-grams, and extract a hash value from each one [14, 15]. The result of this process is a fingerprint that represents the document in each of its sub-parts of length k, further exploited for comparison. Such a procedure, however, does not take into account the behavioral pattern of the plagiarist. In [16], the edit distance is introduced as a similarity metric between chunks of text.

In [17] an hypothetical behavioral pattern of the plagiarist is taken into account. The authors hypothesize that the behavior of the plagiarist consists in the repetition of three prototypical actions: insertion, deletion and substitution. This actions can be performed at any level of the document structure, phrase, paragraph or chapter. Distance between documents is then evaluated recursively exploiting the Levensthein edit distance [18].

All this approaches take into account plagiarism as an operation on text to be considered a mere sequence of characters, with no attempt to capture the likely semantic nature of plagiarism.

The main limit of plagiarism detection algorithms, as a means for identifying similar inventions and patent infringements, is their focus on the language of the description instead of the structure of the invention. Still some lessons learned can be readapted to the specific situation.

In facts, an acknowledged measure of similarity is expressed in the form

$$SIM_{ij} = \frac{keywords_{ij} + keywords_{ji}}{keywords_i + keywords_j} \quad (1)$$

where $keywords_{ij}$ is the number of occurrences of keywords of the document $i$ found in the document $j$ and $keywords_i$ is the overall number of keywords extracted from the document $i$.

An exemplary attempt to reuse in a novel form such a typical plagiarism assessment metric is proposed in [4], where the authors measure patent similarity

by comparing the number of shared SAO triples (Subject Action Object), instead of the keywords alone. The main advantage of the SAO-based approach is that patents are compared in terms of functions delivered by the elements of the invention and general terms are filtered out.

Nevertheless, we observe that while taking into account syntactical information this comparison is still too dependent on the mere text and, as such, it depends more on the writer's style than on the actual 'semantics' of the described invention.

In this paper we propose an alternative approach which is not based on text comparison but on the comparison of the structural and functional architecture of the invention disclosed in a patent.

## 3. A new approach to measure patent similarity

As discussed above, the main limit of the traditional techniques for estimating the conceptual distance between two patents is the dependence on the language style of the inventor.

In order to clarify this concept let's consider the following excerpts:

- US4,713,635: "For example, the barrier portion or insert 107 includes a rib or tongue 109 that is aligned with rib or tongue portions 111, 113, and 115."
- US4,056,798: "One end of the cradle 48 forms a tongue member 50 which is releasably secured within an apertured latch 52 of the trip mechanism 42. […] This deflection causes the bimetal element to engage a hook-shaped projection 66 of the latch 52, pulling the latch 52 to the right and causing the tongue 50 of the cradle 48 to be disengaged from the latch 52."

In both patents, a *tongue member* is a feature of the disclosed invention and can be considered as a subsystem of a further element of the invention (the *barrier portion 107* and the *cradle 48* respectively). The property of being a subsystem is expressed by means of totally different locutions: <component i> "includes" <component j> and <component i> "of the" <component j>. It is worth to notice that the adoption of a SAO-based comparison criterion does not allow to identify this kind of similarity, whatever is the richness and quality of its synonyms list. Similar remarks can be applied also to functional and positional interactions.

In the present paper the authors suggest to evaluate the similarity between two patents by comparing their functional tree [19], i.e. the hierarchical architecture of the invention's components and their functional interactions. In facts, working with the functional tree allows to identify conceptual similarities like the example presented above and to limit the influence of the language style. Moreover, the algorithm described hereafter allows to focus the comparison on a subset of components and/or interactions according to the peculiarity score proposed in [20].

## *3.1 Previous works: automatic functional analysis of patents and extraction of invention peculiarities*

The authors are working on the development of new techniques and algorithms for patent analysis and comparison [19-23]. As a result of these previous experiences a prototype software system (named PatAnalyzer) has been developed with the following functionalities:

- identify the components of the invention;
- classify the identified components in terms of detail/abstraction level and their compositional relationships in terms of supersystem/subsystem links;
- identify positional and functional interactions between the components both internal and external to the system;
- build a thesaurus of "alternative denominations" of the functional elements identified in a given set of patents (hereafter called *project*);
- identify the most relevant components of each patent for a given *project* according to a ranking criterion which combines the detail level of the description with the *Inverse Document Frequency*, i.e. the "rarity" of each synset of the Thesaurus.

In Figure 1 the exemplary results related to the patent US6,064,024 are shown: the conceptual map visualizes the components of the inventions, their hierarchical and functional interactions, as well as their relevance score by means of a color code.

It is worth to notice that the score assigned to the components of each invention allows to select a subset of sentences from the description and the claims of the patent where the top-ranked components are mentioned. In [20] it was demonstrated that such a subset of sentences is sufficient for a "person skilled in the art" for understanding what the core of the patent is about.

In this paper, the top-ranked components and their hierarchical and functional relationships are adopted as a means to compare the inventions of a given *project* in order to estimate their similarity, as described in the following paragraph.

## *3.2 Comparing the functional tree of two inventions*

In this work it is assumed that two technical systems belonging to the same field of application, sharing the same components, structured with the same architecture and characterized by the same functional interactions are conceptually identical. As a consequence, the similarity between two patents is estimated by comparing their components, hierarchical relationships and functions.

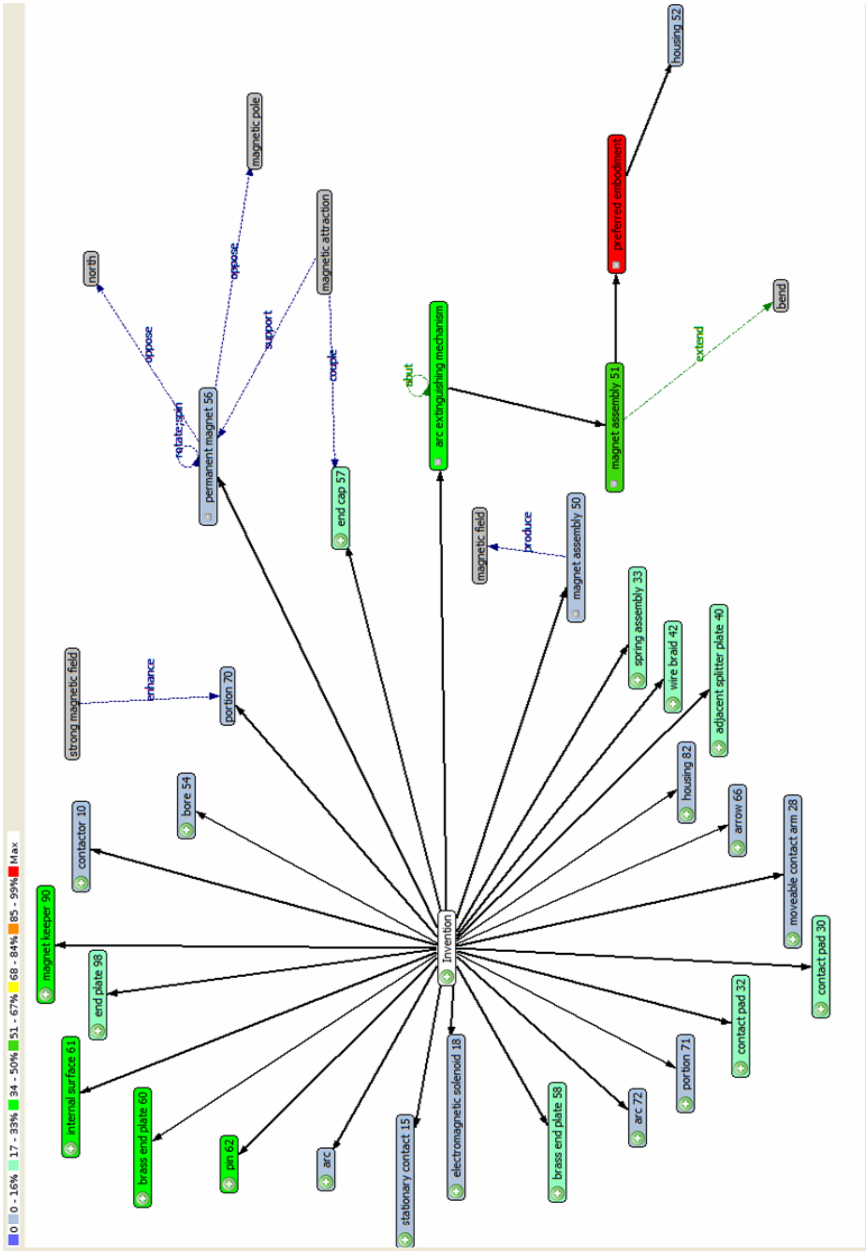



Figure 1. Portion of the conceptual map of the patent US6,064,024: the arrows with no label represent hierarchical relationships (the component at the tip of the arrow is a subsystem of the components at the tail of the arrow); labeled arrows represent functional and positional interactions; the colors highlight the relevance score of each component.

Such a comparison is made also taking into account the alternative denominations of each component, by means of the Thesaurus built according to the rules defined in [20]. More precisely, while comparing the functional trees of two inventions, two nodes are considered equivalent if they belong to the same synset in the Thesaurus of the *project*.

Then the following formula is applied:

$$SIM_{ij} = \alpha \frac{|\Gamma(i) \subset \Gamma(j)| + |\Gamma(j) \subset \Gamma(i)|}{|\Gamma(i)| + |\Gamma(j)|} + \beta \frac{|C(i) \subset C(j)| + |C(j) \subset C(i)|}{|C(i)| + |C(j)|} \quad (2)$$

where $\Gamma(i)$ is the set of hierarchical and functional interactions belonging to the $i$-th patent; $\Gamma(i) \subset \Gamma(j)$ stands for the hierarchical and functional interactions of the $i$-th patent appearing also in the functional tree of the $j$-th patent; $C(i)$ is the list of components belonging to the $i$-th patent; $\alpha$ and $\beta$ are coefficients to weight the mutual relevance of interactions and components.

It is worth to note that the formula (2) can be applied to the whole set of components and interactions extracted from each patent or to a subset of top-scored components and their interactions. Thus, three parameters must be arbitrarily set to evaluate the similarity between two patents: $\alpha$, $\beta$ and $\gamma$, where the latter represents the threshold score for components selection ($\gamma = 0$ means that the whole hierarchical/functional tree is considered to estimate the patent similarity, while $\gamma = 1$ means that only the component with the highest score and its interactions are taken into account).

Whatever is the value assumed by $\alpha$, $\beta$ and $\gamma$, the similarity matrix of a given *project*, built in analogy of the incidence matrix proposed in [24], is a symmetric square matrix $N \times N$ ($N$ being the number of documents analyzed in the *project*) constituted by the similarities of each patents' pair. In other words each cell ($i$, $j$) contains the estimated similarity among the $i$-th and the $j$-th patent.

The rules to define the most suitable values for $\alpha$, $\beta$ and $\gamma$ are still under validation; nevertheless some general directions have already been developed and are briefly discussed in section 5.

## 4. Exemplary application: electrical circuit interruption devices

In order to clarify the procedure described in section 3.2 and to demonstrate its capabilities this chapter reports an exemplary application in the field of electrical circuit interruption devices.

On the base of a previous experience with ABB SACE (www.abb.com), an evolutionary analysis of electrical circuit breakers has been made at the MTI Lab of the University of Florence. A set of 85 patents (ABB *project*) was selected as a combination of two citation trees, i.e. the patents cited from US6,064,024 and US6,373,016 following backward citations up to three levels from the source

patent. Figure 2 shows a portion of the citation tree related to the former patent: each arrow represents a backward citation.

In order to perform such an evolutionary analysis each patent has been analyzed by a technician and a summary has been extracted in the form Problem-Solution (figure 2, close up). In other words, the citation trees have been manually translated into a Problem-Solution tree which can be used with similar purposes of the OTSM-TRIZ Problem Flow Network proposed in [25].

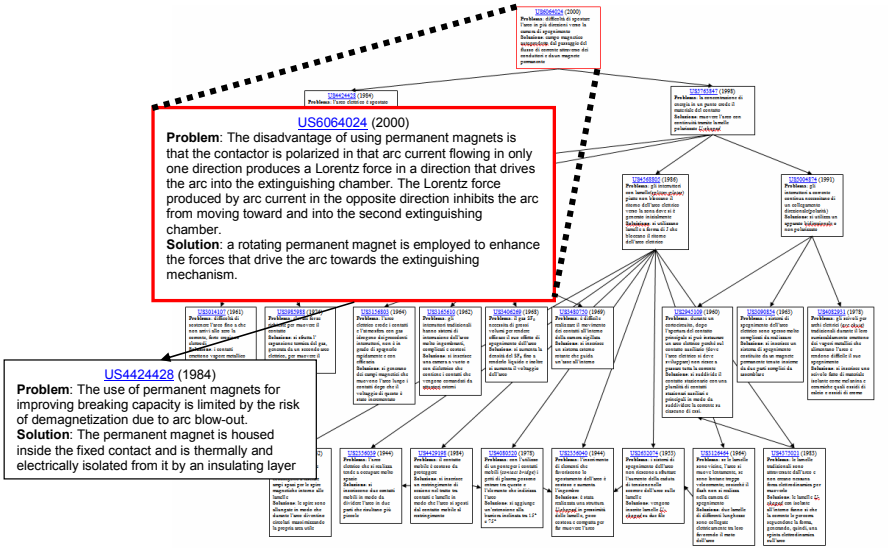



Figure 2. Excerpt from the Patent Citation tree and close up on the Problem-Solution notations associated to a pair of citing-cited patents.

The details of this evolutionary analysis are outside the scopes of the present paper and will be presented in a next publication. Besides, such an extensive analysis carefully operated by humans constitutes a valuable resource to validate the proposed similarity score.

According to the above mentioned procedure, the 85 patents have been processed by PatAnalyzer, thus producing a corresponding number of conceptual maps (as the functional tree shown in figure 1) and related relevance scores. Moreover the analysis has generated a thesaurus specifically dedicated to the ABB *project* (figure 3). The similarity matrix (figure 4) has been calculated assigning the same weight to components and interactions (i.e. $\alpha = 1$, $\beta = 1$) and discarding those components with a score lower than one fifth of the maximum score of the entire project (i.e. $\gamma = 0.2$). The whole computation (functional tree identification, thesaurus generation, definition of the similarity matrix) has required less than 60 minutes on a standard laptop.

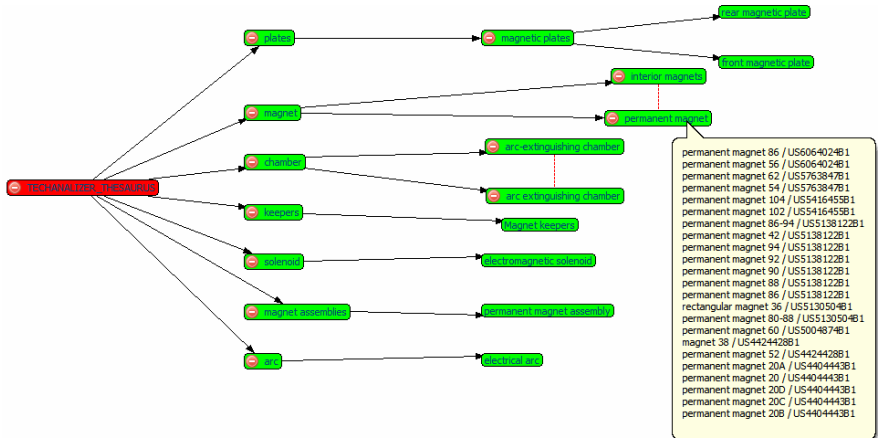
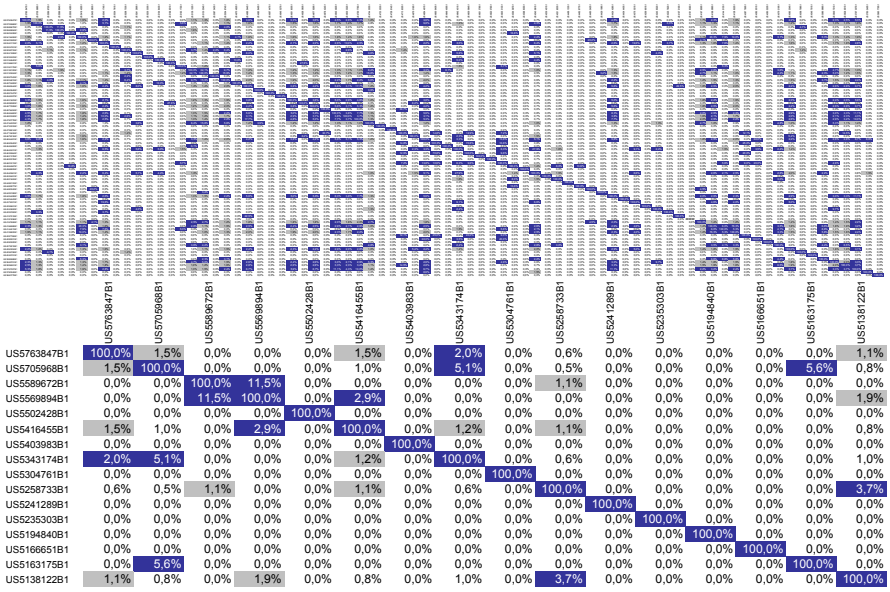



Figure 3. Portion of the Thesaurus graph of the ABB project: the arrows point from hypernyms to hyponyms, (red) dashed lines represent alternative denominations. The box in the right-low corner lists the patents containing the selected word/multi-word.



| | US5763847B1 | US5705968B1 | US5569672B1 | US5569894B1 | US5502428B1 | US5416455B1 | US5403983B1 | US5343174B1 | US5304761B1 | US5258733B1 | US5241289B1 | US5235303B1 | US5194840B1 | US5166651B1 | US5163175B1 | US5138122B1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| US5763847B1 | 100,0% | 1,5% | 0,0% | 0,0% | 0,0% | 1,5% | 0,0% | 2,0% | 0,0% | 0,6% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 1,1% |
| US5705968B1 | 1,5% | 100,0% | 0,0% | 0,0% | 0,0% | 1,0% | 0,0% | 5,1% | 0,0% | 0,5% | 0,0% | 0,0% | 0,0% | 0,0% | 5,6% | 0,8% |
| US5569672B1 | 0,0% | 0,0% | 100,0% | 11,5% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 1,1% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| US5569894B1 | 0,0% | 0,0% | 11,5% | 100,0% | 0,0% | 2,9% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 1,9% |
| US5502428B1 | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| US5416455B1 | 1,5% | 1,0% | 0,0% | 2,9% | 0,0% | 100,0% | 0,0% | 1,2% | 0,0% | 1,1% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,8% |
| US5403983B1 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| US5343174B1 | 2,0% | 5,1% | 0,0% | 0,0% | 0,0% | 1,2% | 0,0% | 100,0% | 0,0% | 0,6% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 1,0% |
| US5304761B1 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| US5258733B1 | 0,6% | 0,5% | 1,1% | 0,0% | 0,0% | 1,1% | 0,0% | 0,6% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 3,7% |
| US5241289B1 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| US5235303B1 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| US5194840B1 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% | 0,0% |
| US5166651B1 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% | 0,0% |
| US5163175B1 | 0,0% | 5,6% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% | 0,0% |
| US5138122B1 | 1,1% | 0,8% | 0,0% | 1,9% | 0,0% | 0,8% | 0,0% | 1,0% | 0,0% | 3,7% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 100,0% |

Figure 4. Similarity matrix for the 85 patents of the ABB project: entire matrix (above) and close up on a reduced number of patents (below). The cells of the matrix can be colored according to one or more thresholds defined by the user (here 1% and 2% respectively) in order to highlight similar patents and clusters.

The resulting similarities have been compared with the outputs of the analysis manually performed by two operators as described above. Such a comparison has revealed a consistent coherence among the two set of results.

As an exemplary demonstration the following patents have revealed relevant matched features: US6,064,024, US5,763,847, US5,130,504, US4,424,428. Indeed all those inventions are characterized by the adoption of a permanent magnet aimed at the deviation and elongation of the electric arc (see also figure 2):

- US6,064,024: "[…] Thus the permanent magnet's strong field will always be oriented to enhance the potentially weak self magnetic field as described with respect to the embodiment in FIG. 1. Therefore the resultant Lorentz force acting on the arc will always be strong enough to drive the arc off the contact pads 30 and 32 and along stationary contact 17 even when the self magnetic field is weak (low current) […]".
- US5,763,847: "[…] As the arc travels into the arc extinguishing chamber 34, it also interacts with the individual magnetic fields produced by permanent magnet 54 in each of the first type of splitter plates 40. […]. The interaction of the arc current with this magnetic field around each plate causes the arc 77 to move in circles on the surface of the splitter plate casing 44. Thus the arc energy is not constricted to one spot on the casing surface as occurred in previous arc chambers, thus erosive effects of arcs impinging the splitter plates are reduced in the present design. […]".
- US5,130,504: "[…] The permanent magnets 80-88 are polarized across the width thereof to establish a magnetic field B (FIGS. 10 and 11) directed front-to-rear through the respective arc chambers, the plates 54 and 90 forming a magnetic path around the outside of the switching apparatus and an air gap across the respective arc extinguishing chambers. […]".
- US4,424,428: "[…]The magnetic field of magnet 38, which is present when the arc appears, leads to rotation of the arc along the annular tracks formed by contact surfaces 34, 36 and rapid extinction of the arc in a well known manner […]".

A selection of the paragraphs containing the components and interactions contributing to the similarity score (in this case "permanent magnet" and related denominations like "interior magnet", figure 3) has been judged sufficient for a person skilled in the art to understand the role of the component and to assess the originality of the solution.

## 5. Discussion and conclusions

The technique proposed in this paper defines the similarity of two patents as the match in terms of functional structure of the inventions, instead of the traditional frequency of keywords co-occurrences. By doing so, patents are grouped into more appropriate conceptual classes and more intrinsically homogeneous clusters can be produced. As explained before, keywords co-occurrence analysis deals with the patents as a whole and considers only the frequency of co-citations. Thus, the result of grouping may be superficial or even spurious since those statistics would not reveal the internal structural relationships between patents.

Besides the proposed algorithm allows to find analogies between inventions described with totally different locutions, while general poorly informative elements (according to a ranking which depends on the specific project and not to a general terminology classification) are neglected.

An open issue is the definition of the rules to assign a proper value to the weights $\alpha$ (interactions), $\beta$ (components) and the threshold $\gamma$. According to the analyses performed so far, the components part of the formula (2) can lead to wrong estimations of the similarity when dealing with a patent having a reduced number of components: in these cases the similarity score is zero when the relevance score of the components is low, since there are no opportunities for matching other patents. Vice versa, if the relevance score of the components of an invention characterized by a reduced number of elements is high, the patent will result highly similar with many patents of the project. Besides, the similarity between patents with a reduced number of components is more suitably assessed by the interactions part of the formula (2). Inversely, in case of inventions with a high number of components described in the patent, also the components part of (2) significantly contributes to the similarity assessment.

A further emerging note is that the contribution of the interactions to the overall similarity score inversely depends on the value assigned to $\gamma$, i.e. the relevance/peculiarity threshold defining the number of components to be considered from each patent in order to perform the comparison. In other words, hierarchical and functional interactions between components provide relevant contributions for similarity assessment if a wider portion of the functional tree is considered for each patent under evaluation, while in a selection limited to the top-score elements from each patent the similarity is mostly evaluated in terms of components.

The authors are involved in a more extensive validation of the proposed algorithm with the aim of providing more detailed guidelines for the definition of the most suitable parameters $\alpha$, $\beta$ and $\gamma$ for a given set of patents.

## References

1. Fluck J., Zimmermann M., Kurapkat G., Hofmann M.: Information extraction technologies for the life science industry. Drug Discovery Today: Technologies, vol. 2, Issue 3, pp. 217-224 (2005).
2. Yoon B., Park Y.: A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. Technological Forecasting and Social Change, vol. 72, Issue 2, pp. 145-160 (2005).
3. Fabry B., Ernst H, Langholz J., Köster M.: Patent portfolio analysis as a useful tool for identifying R&D and business opportunities—an empirical application in the nutrition and health industry. World Patent Information, vol. 28, Issue 3, pp. 215-225 (2006).
4. Bergmann I., Butzke D., Walter L., Fuerste J.P., Moherle M. G., Erdmann V. A.: Evaluating The Risk of Patent Infringement By Means of Semantic Patent Analysis: The Case of DNA-Chips, Proceedings of the R&D Management Conference, 4-6 July 2007, Bremen, Germany, ISBN: 0-9549916-9-9 (2007).
5. Daim T. U., Rueda G., Martin H., Gerdsri P.: Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting and Social Change, vol. 73, Issue 8, pp. 981-1012 (2006).
6. Trappey A. J. C., Hsua F.C., Trappey C. V., Lin C.: Development of a patent document classification and search platform using a back-propagation network. Expert Systems with Applications, vol. 31, Issue 4, pp. 755-765 (2006).

7. Clough P.: Plagiarism in natural and programming languages: An overview of current tools and technologies. Internal Report CS-00-05, University of Sheffield (2000). Available at http://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf. Last access 27 Apr 2008.
8. EVE Plagiarism Detection System. http://www.canexus.com. Last access 27 Apr 2008.
9. Turnitin. http://www.turnitin.com/static/home.html. Last access 27 Apr 2008.
10. Barrett R., Malcolm J., Lyon C.: Are we ready for large scale use of plagiarism detection tools? Proceedings of the 4th Annual LTSN-ICS Conference, NUI Galway, pp. 79-84 (2003).
11. Shivakumar N.: Detecting digital copyright violations on the internet. PhD thesis Stanford University (1999). Available at http://infolab.stanford.edu/~shiva/thesis.html, Last access 27 Apr 2008.
12. Lopresti D.: A comparison of text-based methods for detecting duplication in document image databases. Proceedings of Document Recognition and Retrieval VII (IS and T SPIE Electronic Imaging) San Jose (USA),pp. 210–221, January (2000).
13. Schleimer A. A. S., Wilkerson D.S., Aiken A.: Winnowing: local algorithms for document fingerprinting. Proceedings of the 2003 ACM SIGMOD International Conference on Management of data. ACM 1-58113-634-X/03/06. (2003). Available at http://theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf. Last access 27 Apr 2008.
14. Heintze N.: Scalable document fingerprinting. In Proceedings of the 1996 USENIX Workshop on Electronic Commerce, pp. 191-200 (1996). Available at http://citeseer.ist.psu.edu/348631.html. Last access 27 Apr 2008.
15. Koala Document Fingerprinting (KDF). http://www-2.cs.cmu.edu/afs/cs/user/nch/www/koala-info.html. Last access 27 Apr 2008.
16. Mandreoli P. F., Martoglia R.: Un metodo per il riconoscimento di duplicati in collezioni di documenti. Proceedings of the Eleventh Italian Symposium on Advanced Database Systems, SEBD. (2003).
17. Zini M., Fabbri M., Moneglia M., Panunzi A.: Plagiarism Detection Through Multilevel Text Comparison. In Proceedings of the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution table of contents, pp. 181-185, ISBN:0-7695-2625-X (2006).
18. Levenshtein V.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics-Report, SEBD(10):707–710 (1966).
19. Cascini G.: System and Method for performing functional analyses making use of a plurality of inputs. Patent Application 02425149.8, European Patent Office, 14.3.2002, International Publication Number WO 03/077154 A2 (18 September 2003).
20. Cascini G., Russo D., Zini M.: Computer-Aided Patent Analysis: finding invention peculiarities. Proceedings of the 2nd IFIP Working Conference on Computer Aided Innovation, Brighton (MI), USA, 8-9 October, 2007, in Trends in Computer-Aided Innovation, Springer, pp. 167-178, ISBN 978-0-387-75455-0 (2007).
21. Cascini G., Neri F.: Natural Language Processing for patents analysis and classification. Proceedings of the TRIZ Future 4th World Conference, 3-5 November 2004, Florence, Firenze University Press, ISBN 88-8453-221-3 (2004).
22. Cascini G., Agili A., Zini M.: Building a patents small-world network as a tool for Computer-Aided Innovation. Proceedings of the 1st IFIP Working Conference on Computer Aided Innovation, Ulm, Germany, November 14-15 (2005).
23. Cascini G., Russo D.: Computer-Aided analysis of patents and search for TRIZ contradictions. International Journal of Product Development, Special Issue: Creativity and Innovation Employing TRIZ, vol. 4(1-2) (2007).
24. Yoon B., Park Y.: A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research, vol. 15, Issue 1, pp. 37-50 (2004).
25. Khomenko N., De Guio R., Lelait L., Kaikov I.: A Framework for OTSM-TRIZ Based Computer Support to be used in Complex Problem Management. International Journal of Computer Application in Technology (IJCAT), vol.30 Issue 1-2 (2007).