Abstract
The availability of huge amount of data from a bewildering variety of sources leads to the well-identified nowadays paradox: an overload of information means no usable knowledge. Text Mining and Knowledge Management technologies are therefore assuming a key role for many organizations: in order to propose competitive products or services it is necessary to minimize the resources dedicated to the accomplishment of repetitive tasks and to focus on “creative” activities. Moreover, innovation is basically limited by psychological inertia on one hand, and by lacks of knowledge on the other. This paper presents a set of tools developed by the authors for increasing knowledge extraction from patents. A proposal for a Patent Mining suite is therefore discussed.

Keywords: Natural Language Processing, Functional Analysis, Patent Analysis, Text Mining, Clustering

1. Introduction

Analyses of emerging technologies and their implications are vital to today’s economies, societies, and companies. Such analyses inform critical choices ranging from the multinational level to the individual organization. Decisions that need to be well-informed concern setting priorities for research and development efforts, understanding and managing the risks of technological innovation, exploiting intellectual property, and enhancing technological competitiveness of products, processes, and services [1].

In this context, information retrieval, documents classification, business intelligence, technology forecasting, competitors monitoring etc. are crucial activities requiring advanced tools capable to face the dramatic paradox: an overload of information means no usable knowledge. Such a contradiction between the width of the information source and the low usability of a large amount of documents is typically met in patent search activities: monitoring competitors, checking the novelty of an invention or looking for technical solutions in other fields of application require big efforts even for skilled researchers. Performing a text analysis for constructing design representation has been approached by
several authors, as in [2], by means of techniques mainly based on statistics (i.e. counting
terms frequencies, identifying specific words etc.). The same approach is followed by tools
specifically dedicated to patents analyses as [3], but their main limit is that they cannot
distinguish the role of a component in a technical system.

The commercial availability of software capable of analyzing documents with semantic
processing algorithms offers a revolutionary way to search, summarize and classify
information. Linguistic Analysis tools allow the identification of the key elements of a
document, by combining morphological, syntactic and semantic analyses.

Major results can be obtained by processing structured documents whose format is strictly
related to the content. This characteristic is typical of several forms of technical
documentation, and allows “low-cost” linguistic analysis techniques and tools to be adopted.
Patents are a typical example, with distinguished sections for claims, background of the
invention, description of the preferred embodiment etc.

In this context the authors have developed several tools for supporting any kind of patent
analysis, from detailed searches of inventions peculiarities [4-6] to multi-language patents
unsupervised classification [7-9].

In this paper, first available technologies for patent mining are briefly surveyed; then the
functionalities of the systems developed so far by the authors are summarized. Eventually,
the last chapter presents a work still in progress about the implementation of an integrated
suite for multi level patent analyses and classification.

2. Patent analysis: available technologies

Patent analysis has applications to many different tactical and strategic company
assessments, from those dealing directly with patents, such as intellectual property
management, to those related to company valuation and competitive intelligence. Companies
gain a strategic advantage over their competitors through “technology watch” activities,
which provide the best way of keeping updated with cutting-edge technology developments.
Information can be drawn from patent databases, which can be precious sources of technical
knowledge. Nevertheless, these databases are large and complex and cannot simply be
“watched”. In fact, each patent is identified by specific codes that describe its application
areas, inventor and similar data, as well as by other free textual fields, which are rarely used
for classification purposes. The alphanumerical codes are always partially overlapping and
redundant, the free textual fields contain instead the true valuable information. Thus, it is not
easy, even for an experienced researcher, to recognize the importance of a patent and its
relationship with other patents, especially when the corpus consists of hundreds of
documents, potentially heterogeneous for language used.

Commercially available patent databases provide basic means for information retrieval
and citations tracking, but patents searches are still time consuming and require big efforts
for being accomplished. In facts, citation analyses are the most used techniques for
identifying within a company’s patent portfolio the small number of valuable, high-impact
patents against the large number of patents of marginal importance [10]. It is believed that a
statistical analysis of the rate of publication of patents pertaining to a certain field or assigned
to a certain company, provides information about technology maturity and corporate
technology strategies. Typically, the analysis is performed by counting in an online database
the number of patents issued annually in a set of calendar years [11]. Besides, it normally
takes five or more years from publication before a patent begins to be cited to any great
extent. In general, 70% of all patents are either never cited, or cited only once or twice, so that even five citations place a patent in the top few percent of cited patents [12].

Therefore, the analysis of the free textual description is assuming a greater relevance for getting major advantages from disclosed inventions.

Text Mining applications provide effective means for content searches in the textual fields of patent databases, but they are typically not tailored for patent analyses and too often require a deep expertise about how to gain major advantages from this technology. Some special features are available in Invention Machine Goldfire platform [13] mainly related to the application of syntactic parsing capabilities so that from each sentence a SAO triad (Subject, Action, Object) is extracted. Indeed, such an analysis allows a much more powerful classification of the concepts contained in a patent description; nevertheless, as well as for more traditional keywords based tools, no systems are available on the market for capturing the role of a component in an invention or for grouping patents according to the claimed functionalities apart from their fields of application.

More specifically the following features aimed at speeding-up patent analysts activity still lack on the market:

– identifying the architecture of the claimed invention, distinguishing the functional (semantic) role of each component;
– identifying invention peculiarities as a means for providing an automatic extraction of the core of the patent; (it is worth to mention that too often the patent abstract is very low informative);
– clustering technical solutions according to the way a function is accomplished apart from the field of application (therefore providing proper means for technology transfer);
– allowing easy and effective queries by means of a multi-language taxonomic knowledge base so that search results do not depend on patent language and/or the use of synonyms, hyperonyms, meronyms etc.

3. Tools for efficient patent analyses

On the base of the above discussed limitations of traditional tools for patent analyses, the authors have carried on complementary experiences that can be integrated with the aim of providing a comprehensive platform for any kind of intellectual property related tasks.

In this chapter a survey of these works is presented.

3.1 Automatic patent functional analysis

Functional analysis is a powerful tool for conceptual design both for problem identification and innovative solutions generation: the functional description of a product is a description at an abstract level, so that different design solutions can be explored by developing functional variants. At the same time, functional analysis helps the designer in following a systematic approach also in the study of complex systems, by breaking up functions into simpler sub-functions and subdividing the problem into more manageable tasks.

According to these assumptions in [4, 5] a system for translating automatically the description of an invention into a functional diagram has been proposed. The methodology has been implemented in a prototype software tool, PAT-Analyzer, capable to identify the components of the patented system, perform a hierarchical classification of those components subdividing them in different abstraction levels, draw a functional diagram of the complete system and of the detailed subassemblies, therefore providing a graphical
representation of the invention described in a patent. An exemplary diagram obtained by processing the US Patent 5,406,868 “Open End Wrench” is shown in Figure 1, where:

1) each identified component of the system is represented by a rectangle labeled with its reference number and the representative name defined in the Components Recognition phase; each identified component or subject external to the system is represented by a grey rectangle labeled with a representative name;

2) the detail level hierarchy is represented nesting the components at a deeper detail level inside the components at a more abstract level;

3) the functional interactions between the identified components are represented with straight arrows pointing from the Tool to the Artifact, labeled with the Field;

4) the positional interactions between the identified components are represented with dashed arrows pointing from the Tool to the Artifact, labeled with the Field.

By means of several types of score ranks it is possible to highlight the core of the invention, the most peculiar components and performed functions:

1) Detail Level Chart: on the basis of the components hierarchical classification performed in the text analysis phase, it is possible to assign a Detail Level (DL) to each TFA triad and/or to each paragraph: a DL is assigned to each component so that the maximum abstraction level is represented by a DL=0 and the DL of each subsystem is one level greater than the DL of the corresponding supersystem. The detail level of a basic sentence (TFA) is estimated as the average DL of Tool and Artifact; the detail level of a paragraph is the average DL of all the TFAs belonging to that paragraph. By analyzing the DL run along the description it is easy to identify the most relevant paragraphs and concepts of the inventions (it is worth to notice that if the patent description is focused on specific details of the proposed system, it means that such a part of the description is strictly related to the peculiarities of the invention).

Figure 1: Functional diagram extracted from the US Patent 5,406,868 ‘Open End Wrench’. The color map highlights the sub-functions accomplished at a more detailed level.
2) Components Occurrence Analysis: by counting the citations of each component of the proposed invention (of course taking into account of the automatically extracted synonymous) it is possible to determine a relevance score of the components themselves; in order to improve the quality of such an estimation, different weights are assigned to the citations found in the title, abstract, independent claims, dependent claims (first level, second level, others), summary, description of the preferred embodiment.

3) TFA Occurrence Analysis: by means of the same technique adopted for the Component Recurrence Analysis, it is useful to assign a score to the functional and positional triads TFA. Also partial pairs of the triads (TF, FA, TA) are counted by means of the same set of weights. Therefore a rank of the most relevant sub-functions performed by the invention is determined, again in order to focus the attention of the user to the patent peculiarities.

3.2 Concepts mapping

The results of a patent analysis accomplished by means of the tool mentioned in the previous paragraph, can be further processed in order to extract a contents map of the patent itself. In facts, if the above described analysis allow to identify the core topics of the invention, for example by looking for the paragraph with the highest detail level of the description, further useful information can be found also in other sentences with a lower detail score, but still treating the same concepts.

By means of a TFA-based content analysis as described in [6], it is possible to create a map of the concepts disclosed in a patent. Basically, the technique consists in calculating for a given TFA identified as relevant for the examined invention, a TFA score for each paragraph, by weighting differently the occurrences of the Tool, the Field and the Artifact.

An exemplary application of this technique is shown in Figure 2, where the three top score concepts in terms of detail level of the US Patent 6,097,012 (Induction-heating bender) are mapped through all the patent text.

![Figure 2: Content Analysis of the top score concepts identified by the analysis of the US Patent 6,097,012: heating and cooling system for non-uniform temperature distribution in the beam section (top score); rotating gripper head for variable bending radius (2nd top score); feeding mechanism (3rd top score).](image-url)
Tab. 1: Summary of the US Patent 6,097,012 obtained by collecting the top score paragraphs, according to the analysis shown in Figure 2. The paragraph score is evaluated by summing the content score of the three more relevant TFAs.

<table>
<thead>
<tr>
<th>Par. #</th>
<th>Text</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>An induction-heating bender according to claim 1 wherein the heating-and-cooling mechanism comprises a heating coil encircling the workpiece to be bent and a plurality of cool compartments placed inside of the heating coil.</td>
<td>3.5</td>
</tr>
<tr>
<td>13</td>
<td>An induction-heating bender according to claim 6 wherein the heating-and-cooling mechanism further comprises at least one circular aperture and at least one elongate slot to cool the temperatures of the plurality of cool compartments independently.</td>
<td>7</td>
</tr>
<tr>
<td>28</td>
<td>As described above, the gripper 21 is rotatably fixed to the free end of the pivot arm 20 for gripping the workpiece M. At the opposite end, the pivot arm 20 is rotatably fixed to the square arm support 30 via an associated arm holder 37, as later described.</td>
<td>2.3</td>
</tr>
<tr>
<td>30</td>
<td>The pivot arm 20 is inserted in the square holder 37 by loosening fastening bolts 37a, and then, it can be fastened to the square holder 37 by tightening the fastening bolts 37a, thereby permitting the pivot arm to be set at a desired pivot-to-gripper distance. As shown in the drawing, the upright axle or pivot 36 of the square holder 37 is offset from the longitudinal axis of the pivot arm 20 horizontally in the direction in which the workpiece travels.</td>
<td>3.5</td>
</tr>
<tr>
<td>36</td>
<td>These coolant ejection nozzles 48 can eject different coolants at different flow rates, as follows: water may be ejected from one of these coolant ejection nozzles 48 at the temperature of 40°C, and at the same time, air may be ejected from the other coolant ejection nozzle 48 below the temperature of 40°C, preventing the hardening of the workpiece.</td>
<td>7.8</td>
</tr>
<tr>
<td>37</td>
<td>Referring to FIGS. 6A to 6D, coolant slots may be formed in place of the ejection nozzles 48 as shown in FIG. 5. These coolant slots are of different shapes and sizes for controlling the flow rate of the coolant. A linear arrangement of relatively small circular apertures 48a is appropriate for ejecting the coolant at a decreased flow rate (see FIG. 6A). A linear arrangement of elongated slots 48b is appropriate for ejecting the coolant at an increased flow rate (see FIG. 6B). An alternate, linear arrangement of circular apertures and elongated slots may be used for ejecting the coolant at controlled flow rate (see FIGS. 6C and D).</td>
<td>4.7</td>
</tr>
<tr>
<td>38</td>
<td>Referring to FIG. 4 again, the divisional sections 42i, 42f and 42h of the heating-and-cooling coil sub-unit 41 for heating and cooling the inner flange of the H-steel have circular apertures 48a and elongated slots 48b made for rapid cooling whereas the divisional sections 42a, 42b and 42c of the heating-and-cooling coil sub-unit 41 for heating and cooling the outer flange of the H-steel have circular apertures 48a made for slow cooling. The cooling of each flange at different temperatures permits the wrinkle-free bending of the H-steel.</td>
<td>2.2</td>
</tr>
<tr>
<td>43</td>
<td>Referring to FIG. 10, the manner in which the subsequent two meter-long section is bent at two meter-long radial distance is described. First, the pivot arm 20 is returned to the initial position in which the pivot arm 20 is parallel to the arm support 30, and the fastening bolts 37a of the square holder 37 are loosened. Then, the motor 34 is made to start rotating, thereby moving the carrier 33 two meters apart from the center of the heating-and-cooling mechanism 41. The turntable 25 of the gripper 21 is inclined at the angle (\alpha) with respect to the longitudinal axis of the pivot arm 20 to be in conformity with the precedent, one meter-long section already bent (see FIG. 3). The fastening bolts 37a of the square holder 37 are tightened to hold firmly the pivot arm 20. While the workpiece is heated by the heating-and-cooling mechanism 41, the beam-like workpiece is fed forward by the conveyor 11, thereby allowing the beam-like workpiece to be bent. The turning of the turntable 25 in conformity with the precedent curved section of workpiece assures the smooth transition from the precedent to subsequent curved sections, which smooth transition cannot be attained simply by bending at different radial distances as is the case with the conventional bender.</td>
<td>2.7</td>
</tr>
</tbody>
</table>
By selecting the paragraphs with a score greater than a threshold value it is possible to identify a meaningful subset of sentences that constitute an efficient synthesis of the patent content, as shown in Tab. 1. It is worth to note that compared with standard linguistic summarization techniques, such an approach is strictly based on the contents of the invention description, therefore it is very well focused and no irrelevant concepts are included.

3.3 Patent clustering

In the previous two paragraphs the attention has been focused on the detailed examination of a single document; besides, while approaching the analysis of several dozens or even several hundreds patents first it is necessary to classify their content apart from the invention field of application.

As detailed in [7, 8], such a text mining task can be accomplished following a two step process: the Linguistic Analysis of documents allows to recognize the relevant descriptors of the content, whilst the Statistical Analysis uses those descriptors to dynamically discover the thematic groups that best describe the document categories.

The automatic Linguistic Analysis is based on Parsing, Morphological and Statistical rules [14, 15]. The Parsing analysis is based on a set of pre-defined rules, which specify the most relevant fields in patents and their main features: the label identifying the field of interest, or the masks that need to be applied to extract the main information included in the field (i.e. inventor name, assignee name etc.).

The automatic linguistic analysis of free textual fields is based on Morphological and Statistical criteria. This phase is intended to identify only the significant expressions from the whole raw text. This analysis recognizes as relevant terminology only those terms or phrases that comply with a set of pre-defined morphological patterns (i.e.: noun+noun, noun+preposition+noun sequences etc.) and whose frequency exceeds a threshold of significance [16]. The detected terms and phrases are then extracted, reduced to their Part Of Speech tagged base form. Once referred to their language independent entry inside a multilingual dictionary, they are used as descriptors for documents.

Indexation based on terminology detection is extremely reliable for managing any type of documentation, especially if it is technical and scientific.

Once the documents have been linguistically structured, they are clustered into thematic groups by using the Relational Data Analysis (RDA) method [17], whose greater advantage is that the resource consumption (process time and memory) is linear vs. the number of analyzed documents. The clustering process is carried on by linguistic and statistical criteria by using:

- grammatical categories as descriptors for documents (i.e.: proper names are the most valuable items for subject detection, adverbs are obviously not so important);
- semantic categories as descriptors too;
- filters for descriptors, assigning different weights to them and their values;
- an upper and a lower thresholds for descriptors values, so as to avoid considering noisy words (i.e.: an item, whose occurrence is systematically or relatively rare in the corpus, can be discarded, giving no additional information);
- a threshold for similarity.

An absence/presence table, describing the cross connection between the documents and the words they contain, produces the Similarity table: documents sharing lemmatas are then grouped into thematic clusters by the quasi-seriation of this table, each group identifying a
topic, respecting the similarity criterion. The “weighted Condorcet criterion” is normally used to detect similarities among all the documents.

As an example, let’s consider a study performed on more than 200 patents and scientific articles related to airbag and car crash tests. The selected data illustrates the classification problem, being each document characterized by a number of items which describe its contents: title and abstract lemmas, time/date, document type, classification codes which are partly overlapping and partly redundant, due to the fact that documents have been taken from different data banks.

The Text Mining analysis automatically detects nine groups of documents, each with highly distinctive characteristics. The results shown in Figure 3 are similar to what everyone would expect from reading these type of documents: the group of documents which deals with whiplash injuries is strictly linked to documents related to seat characteristics, head restraint, seat belt and cervical spine/vertebra.

It is worth to note that side impact documents, instead, are completely isolated from the others: it is intuitive to think that seat characteristics, head restraint, seat belt have a very low influence in limiting neck injuries in lateral car crashes. Therefore, a simple reading grid is given to the analyst, who is able to access documents by contents and find out any hidden relationship among them.

Figure 3: Automatic classification by clustering more than 200 patents related to whiplash injuries.
### 3.4 Multilingual searches

While big companies are usually familiar with patent searches both for strategic and technical activities, just a small number of small and medium enterprises make use of patent databases for improving their competitive means. Very often a big obstacle is constituted by linguistic barriers, therefore limiting the efficiency of free databases like esp@cenet.

In order to encourage SME participation in patenting and technology transfer opportunities the Wisper project (Worldwide Intelligent Semantic Patent Extraction and Retrieval [9]) has been developed with the aim of:

- providing customized content of publicly held patent data;
- employing techniques of deep level data mining and developing partial ontologies of the patent domain;
- developing techniques for search on figures, drawings, and images, opening up a whole new realm of web-based semantic information in patents;
- supporting Natural Language Information Retrieval, producing results in native language from texts in different languages;
- maintaining a user profile of previous searches, preferred language and areas of interest.

The prototype automatically recognizes sentences written in five natural languages (English, French, German, Italian and Spanish) basing on wide monolingual dictionaries and grammars. Synthema monolingual lexicons have the following sizes:

- English 100,000 lemmas,
- German 70,000 lemmas,
- French 55,000 lemmas,
- Italian 62,000 lemmas,
- Spanish 42,000 lemmas.

The prototype also uses a multilingual lexicon specifically created starting from the English catchwords retrieved from the IPC 7 version. Each English catchword is tagged with class information and translated into the other four supported languages. The translation work, which has started from the B class, is still in progress.

The user can enter a query phrase in his/her own natural language. The prototype performs a morphosyntactic analysis of the query and identifies the user language and relevant terms, then looking for those terms in the local DB shows the query results in the IPC List and relevant catchwords in the homonymous panel (Fig. 4, above).

It is now possible to tune the query in 2 ways: by class or by catchword. The panel IPC List shows the list of relevant IPC codes with their description in the user language. Choosing one or more classes in the IPC List window, it is possible to obtain all the patents of the selected classes from local DB (Fig. 4, middle).

The catchwords panel shows the relevant terms extracted in the user language, the related IPC code and the corresponding terms in all supported languages. To retrieve patents, the interface lets the user check the language he is interested in.

While browsing the results of the query, the user can ask for an automatic translation of the patent into his language (Fig. 4, below); it is clear that the quality of machine translation is not high. Anyway, even if not all sentences may be linguistically correct, the translated abstract may absolutely help a user in understanding the content of the invention.

At the present stage of the project, all the (38) B62B catchwords are correctly recognized in all the five languages. A test performed on other 100 catchwords of the B CLASS revealed the following coverage: all the English catchwords are recognized; about 70% of them are recognized in French and Italian; about 40% of them are recognized in Spanish and German.
Figure 4: Wisper multilingual portal for intelligent access to massive patent databases.

On the basis of the complementary competences experienced by the authors, University of Florence and Synthesma srl have established a collaboration aimed at integrating the patent mining capabilities presented so far.

By combining the functionalities described in the section 3 of the paper, it is possible to implement a platform to be tailored according to specific customers requirements with the following general characteristics:

- multi-language queries for patent retrieval;
- patent classification based on the free textual description of the inventions and identification of not evident conceptual links among patent clusters;
- graphical representation of inventions functional architecture;
- identification of the core concepts disclosed in each patent and pertinent paragraph extraction for patent summarization.

Furthermore new opportunities arise from this integration, once more confirming that a proper collaboration is more effective than a trivial sum of capabilities.

First, the clustering algorithms can be applied not just to the descriptors extracted following the approach presented in par. 3.3, but to the outputs of the functional analysis described in par. 3.1, i.e. to the list of components of an invention or to their functional relationships. Otherwise, the standard descriptors extraction can be applied not to the whole patent text, but just to the most relevant paragraphs identified according to the par.3.2 methodology. Both these original techniques are very promising in terms of capturing inventions peculiarities and semantic relationships. In facts, it is worth to notice that compared with available “semantic” technologies merely based on text syntactic parsing, in this case e real semantic analysis is performed by associating components, functionalities and roles.

The preliminary tests aimed at checking the efficiency of the proposed technique were very encouraging even if accomplished on a restricted number of patents in the field of medical diagnosis x-ray apparatuses. A more careful validation is in progress in the high-tech textile field and will be detailed in a next publication.

Another relevant novelty is constituted by the capability of providing an even rough translation of the summary of an invention, not based on the abstract written by the inventors, but on the paragraphs selected according to their relevance score.

5. Conclusions

In order to face the unmanageable size of technological information sources, means for speeding up reading capabilities are necessary for any product development activity.

It has been proven that the application of linguistic analysis techniques to patent documents allows efficient information retrieval and classification, but still a lot of opportunities for improving patent mining efficiency can be exploited.

In this paper, after surveying commercially available systems, the patent mining capabilities of the technologies developed by the authors have been presented. The complementarities of these experiences have suggested the integration of such functionalities in a patent mining platform capable of providing an useful support to any level patent management activity, both to expert analysts and sporadic users (as for example designers in SMEs).
References


