ELSEVIER

World Patent Information 25 (2003) 335–342

WORLD
PATENT
INFORMATION

www.elsevier.com/locate/worpatin

Text mining applied to patent mapping: a practical business case

Michele Fattori [a,*], Giorgio Pedrazzi [b], Roberta Turra [b]

[a] *Tetra Pak Carton Ambient SpA, via Delfini 1, 41100 Modena, Italy*
[b] *CINECA, via Magnanelli 6/3, 40033 Casalecchio di Reno (Bologna), Italy*

Abstract

Professional patent searchers are traditionally rather suspicious of the alleged "black box" effect inherently attached to intelligent software engines relying upon linguistic technologies for patent analysis and mapping. In this article, the authors propose that such prejudices can be overcome by setting a realistic business objective while experimenting with these new linguistic tools, as well as by applying serious methodology for validating the results of the analysis. The strengths and weaknesses of a particular text mining tool are assessed with reference to a practical business case in the field of packaging technology, and a comparison of the outcome of such an analysis with a traditional one, carried out using conventional patent classifications, is also described.

Keywords: Text mining; Data mining; Patent mapping; Patent analysis; Clustering techniques; Competitive intelligence; Intellectually assigned patent classifications; Results validation; Linguistic technology; Packaging technology

1. Introduction

Various data and text mining tools applied to patent analysis have been around for quite a while now [1,2]. Nevertheless, as pointed out by Krier and Zaccà [3], within the professional patent information community there still is a high degree of scepticism as regards the use of these new linguistic technologies. At least in part, this is due to the relative "black box" effect [1] inherently attached to the nature of the said technology.

Not surprisingly then, professional patent searchers are rather suspicious of tools that do not generally grant the user complete control over their inner workings.

Indeed, an aim of this article is to try to shed some light on the degree of control that a user can expect to experience when carrying out some kinds of patent analysis with the help of text mining techniques.

The adopted point of view is that of a patent information professional, who is not necessarily an expert in linguistic algorithms. The idea was to extrapolate, if not a precise set of rules, at least some useful guidelines that could be applied to a fairly large number of real business scenarios when patent analysis is combined with text mining techniques.

An experimental text mining tool prototype named PackMOLE™ (mining online expert on packaging patents) was jointly developed by the Data Mining Centre of the CINECA (Consorzio Interuniversitario per il Calcolo Automatico dell'Italia Nord Orientale) consortium and the Intellectual Assets Department of Tetra Pak Carton Ambient SpA. In fact, the PackMOLE™ prototype came into being thanks to the combined expertise acquired by CINECA in the field of text mining applications and that of Tetra Pak Carton Ambient SpA in the management of intellectual assets. The original aim of the PackMOLE™ project [2] was to implement an application for mining patent information in the packaging field.

2. The PackMOLE™ prototype

The PackMOLE™ prototype is able to work under unsupervised conditions, which means that the documents (patents, in this case) are processed and grouped

* Corresponding author. Tel.: +39-59-898-009; fax: +39-59-898-027.
E-mail addresses: michele.fattori@tetrapak.com (M. Fattori), g.pedrazzi@cineca.it (G. Pedrazzi), turra@cineca.it (R. Turra).

[1] According to Krier and Zaccà, it is not clear "...whether [that] box is really black or just looks black because of absence of illumination (in this case knowledge of the linguistic algorithms used)".

[2] In the course of this article we shall refer indiscriminately to PackMOLE™ as both the project and the tool.

0172-2190/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0172-2190(03)00113-3

into clusters that are dynamically generated by the algorithm, depending on a number of criteria which are not predetermined on the basis of a user-provided taxonomy.

This particular process is called *clustering*, as opposed to *categorisation* techniques that are dependent upon some kinds of predictive model and often also need adequate training [2].

Categorisation techniques are being investigated by various patent offices for implementing systems for the categorisation and classification of patent documents, and have already been discussed in recent issues of this journal [3–5].

Rather than just classifying documents, clustering techniques can yield valuable insight into the relationships existing between different categories (or clusters) of documents, thus a clustering approach to text mining is considered more effective in a business environment, especially where patent information is regarded not only as a support for legal issues, but also as an important player within the competitive intelligence function.

In this context, the original assumption that formed the basis of the PackMOLE™ project was that text mining or, more precisely, the particular type of text mining called clustering, could not be properly considered a search technique inasmuch as traditional patent searching was concerned. Rather, text mining (or clustering) should allow for the extraction of information regarding *patenting trends* in a more efficient way when compared to the capabilities of the standard Boolean tools which, on the other hand, allow for more precise retrieval of *patent documents* and other bibliographic information regarding patents (such as legal status and patent families).

The outcome of a text mining (or clustering) session primarily consists of a set of clusters, i.e. groups of documents that show a certain amount of similarity, according to a threshold value.

Each cluster is labelled with one or more keywords deemed to be representative of its content, thus it is possible to have a preliminary idea of the said contents without actually reading all the documents.

Of course, it is also possible to browse through the documents contained in each cluster to review the quality of the clustering process, as well as to display a graphical representation of the said clusters. We call this graphical representation the ''bubble map'' (or ''patent map'', since all the documents involved in our tests were patent documents), with the clusters being represented as bubbles [6].

If the mining algorithm finds that there is a certain degree of similarity between different clusters, those clusters are ''linked''. In the patent map, such links are displayed as coloured lines connecting two (or more) bubbles. The colour of the link lines varies according to the relative strength of the link. In fact, the presence of such links is created thanks to a second threshold value, which is automatically determined by the software.

It is also possible to visualise the properties of the various clusters with histograms, as well as exporting a wealth of non-textual, bibliographic cluster data (i.e.: applicants' names, patent classifications, priority and filing dates, etc.) into external software programs for further processing. We shall refer to this kind of patent data with the term *metainformation*.

Among the various clustering algorithms and techniques currently available, the one called *relational analysis* is known to be particularly efficient when processing textual information and was therefore selected as the algorithm of choice to be implemented into the PackMOLE™ prototype.

In our case, the textual information to be analysed consisted of the contents of selected fields of standard Derwent World Patents Index® records retrieved from an in-house intranet database. The Derwent WPI dataset was selected in order to describe the complete patent portfolio of a particular company, for testing the PackMOLE™ prototype against a real business case.

3. Methodology

3.1. The tool

With the PackMOLE™ prototype, the user has the option of customising a number of different parameters that govern the clustering process:

- the maximum number of clusters allowed,
- the weighting system,
- the keyword drop threshold (KDT) and
- the minimum domain homogeneity (*alpha*).

A brief explanation of the meaning of these parameters is given below.

The *maximum number of clusters allowed* parameter has quite a self-explanatory name. Nevertheless, it is worth noting that, when necessary, the software automatically generates a number of clusters that is lower than this specified upper limit.

The *weighting system* can assume three different values: *large domains*, *specific domains* and *medium domains*. By selecting large domains, the algorithm tends to create large clusters based on frequent words; with specific domains, small clusters based on rare words are likely to be created, while medium domains is a compromise between the two.

The *KDT* parameter teaches the algorithm to ignore those words that happen to appear in a lower number of documents than the specified value.

Table 1
Statistical indexes

Index	Description
Frequency	The number of documents, in each cluster, having that particular feature
Characteristic ratio	The percentage of documents, in each cluster, having that particular feature. The higher this measure is, the better that feature characterises the documents of the cluster
Global frequency	The number of documents, considering the whole dataset, having that particular feature
Global ratio	The percentage of documents, considering the whole dataset, having that particular feature
Discriminant ratio	The ratio of frequency to global frequency. The discriminant ratio is the percentage of documents, in each cluster, that have that particular feature, with respect to the number of documents in the whole dataset that also have that particular feature. The higher this measure is, the better that feature discriminates the documents in the cluster from the remaining documents in the dataset

Finally, the alpha parameter sets the minimum degree of similarity that two documents shall possess in order for them to be grouped within the same cluster. [3]

Other than setting values for the mentioned parameters, the user also has the option of selecting one or more stop words. This can be useful for the exclusion of meaningless words from the clustering process [6].

Obviously, some of these parameters can affect the behaviour of the algorithm in contrasting ways and, in some cases, the user has to pay attention to avoid any possible conflict. For instance, by raising the value of the alpha parameter, the algorithm is naturally geared to produce more clusters of a lower size, which could be inconvenient if the maximum number of clusters allowed was set to a low value.

In any event, in extreme cases, the PackMOLE™ tool has the ability to automatically override incongruous user settings.

Other typical clustering parameters or functions, namely the *similarity index* and the *number of iterations*, are not user customisable within the PackMOLE™ environment and therefore not discussed here.

In order to help the user to evaluate, in an objective way, a particular combination of the above mentioned clustering parameters and, therefore, the outcome of the corresponding clustering session, one can rely upon three built-in criteria called

- within,
- between,
- quality.

The *within* criterion (or *intra-cluster homogeneity*) is a measure of the internal homogeneity of each cluster. On the contrary, the *between* criterion (*inter-cluster separability*) measures the degree of similarity between different clusters [7].

The *quality* criterion in some way summarises the two preceding ones.

[3] There is actually a second threshold parameter other than alpha. This second parameter is responsible for the creation of links between different clusters having a similarity degree between the two thresholds.

The PackMOLE™ prototype is able to provide the user with within, between and quality values related to each single cluster as well as to each clustering (or cluster map) as a whole.

Generally speaking, a good clustering is expected to feature high values for its within and quality criteria and a low value for its between criterion. There are some caveats, for instance: a very high "within" value could mean a cluster map consisting of very small clusters (in theory, so small as to comprise just one document), and a very low "between" could well mean that all the clusters in the map are completely disconnected. In the first case, the cluster map is obviously of no value whatsoever, while in the second case the map lacks completely what possibly is the most remarkable value added information that could be obtained through clustering analysis: the indication of relationships or links between clusters.

It is the opinion of the authors that these criteria, though valuable for deeper insight and fine-tuning of the clustering process, as well as for obtaining a first gross indication of the overall quality of a clustering session, should nonetheless be handled with extreme care and, most importantly, should not be regarded as a feasible shortcut for validating a cluster map a priori.

Finally, we note that the available metainformation is automatically processed by the PackMOLE™ tool and, for each of the non-textual features mentioned earlier, a number of statistical indexes are also calculated (see Table 1).

All these criteria and indexes help the user in selecting the most appropriate combination of customisable parameters and thus are an indispensable aid during the fine-tuning and calibration steps of the mining process. Nevertheless, as previously mentioned, during the step of validating the results, the user should rely greatly upon his/her knowledge of both the mining tool and the subject matter of the documents involved.

This validation step can be performed manually, i.e. by reading every document in each cluster, can rely upon some kind of calculated statistical index, be facilitated through the use of a number of different graphical tools, or any combination of the above. In any event, this difficult but necessary step is likely to be rather time-consuming,

especially in the case of large and complicated patent maps.

3.2. Analysis of a patent portfolio: the mining steps

For testing the PackMOLE™ prototype, it was decided to select a realistic application, i.e. one that could easily fit into the patent information process of Tetra Pak Carton Ambient SpA and thus adequate emphasis was on extracting new and actionable knowledge in a real business context. At the same time, the capacity and limits of the tool itself had to be taken into account, to make sure we obtained a valid set of results.

As a first step, we therefore selected all the patents filed by a particular industrial group, active in the packaging and other fields, during the 1991–2000 time range. To enable a better view of the patenting dynamics, we adopted the well-known technique [6] of further splitting the time range into smaller segments or slices: In this case, two segments were created, corresponding to 1991–1995 and 1996–2000.

The first segment was populated by 86 distinct documents, while the second segment was slightly bigger, accounting for 106 documents. As a consequence of the nature of Derwent WPI records, all the documents retrieved were conveniently representative of unique patent families.

A number of experimental tests were conducted in order to find the best possible combination of clustering parameters for the chosen application. In particular, we wanted to obtain the highest possible value for the overall quality criterion, while at the same time keeping under control the values of the within and between criteria.

Of course, we also did not want to obtain a high number of small, meaningless clusters, nor did we want to lose important information about the relationships between different clusters. Therefore, we conducted several tests for evaluating the response of the algorithm with respect to a change in the clustering parameters when mining the two different datasets or time slices.

We set the maximum number of clusters allowed parameter to 30 and the KDT parameter to 5 and let the other parameters vary, as shown in Tables 2 and 3, which refer to the first and second time slices, respectively. Table 4 summarises the meaning of the symbolic values [4] shown in the preceding tables for the alpha and weighting system parameters.

Table 2
Calibration of parameters: first time slice (1991–1995)

Alpha	Weighting system	Quality	Between	Within
0	0	75.24883	20.22216	53.61754
25	0	75.70516	20.4978	55.41937
50	0	76.45833	21.42763	60.51506
75	0	76.71793	21.5146	62.39135
100	0	77.0136	22.28996	68.95846
150	0	77.02991	22.56356	71.63733
200	0	76.97636	22.88898	74.86402
250	0	76.81571	23.1361	75.97598
0	50	77.16235	18.40972	53.2596
25	50	78.78848	18.04413	57.47902
50	50	78.77113	19.8377	64.65002
75	50	79.66386	18.8663	65.21858
100	50	79.00095	20.26706	69.66037
150	50	79.02204	20.47135	71.8409
200	50	78.94551	20.72742	73.73237
250	50	78.90068	20.82032	74.26834
0	100	79.53323	17.48709	57.26816
25	100	80.16075	17.96368	62.31783
50	100	80.37956	18.19908	65.07483
75	100	80.46864	18.42106	67.185
100	100	80.57274	18.60637	69.6718
150	100	80.58289	18.85936	72.1897
200	100	80.54706	19.0326	73.67558
250	100	80.43496	19.23806	74.61258

Table 3
Calibration of parameters: second time slice (1996–2000)

Alpha	Weighting system	Quality	Between	Within
0	0	75.26623	21.37498	53.54976
25	0	75.74541	21.71035	56.20657
50	0	76.2831	22.12541	60.64929
75	0	76.44939	22.28767	62.69138
100	0	76.65265	22.77321	68.25748
150	0	76.63417	23.0164	70.73466
200	0	76.56703	23.21742	72.4791
250	0	76.42722	23.3527	72.05145
0	50	79.61112	18.40944	59.34375
25	50	79.744	18.53449	60.78518
50	50	80.06667	18.84157	65.34303
75	50	80.11666	18.93831	66.59558
100	50	80.18723	19.07035	68.59628
150	50	80.18952	19.3242	71.39236
200	50	80.19743	19.32966	71.59061
250	50	80.12227	19.44516	71.85658
0	100	82.21957	16.05856	61.68178
25	100	82.48489	16.27234	65.33604
50	100	82.56562	16.39861	67.12593
75	100	82.61001	16.44113	68.0085
100	100	82.65516	16.48689	68.98714
150	100	82.65041	16.69681	71.03367
200	100	82.65862	16.74308	71.69827
250	100	82.49071	16.89322	70.78495

It can be easily seen that, at least in terms of the quality criterion and for both segments, the best results were obtained when the weighting system was set to specific domains.

[4] The alpha parameter can assume continuous values within the [0, 1] interval, while the weighting system parameter can only assume three discrete values: large domains, specific domains, or medium domains. To consistently compare these two parameters, as shown in Tables 2 and 3, against the three built-in criteria (quality, between and within), the use of a graduated, symbolic scale (shown in Table 4) was found to be appropriate.

Table 4
Explanation of symbolic values shown in Tables 2 and 3

Symbolic value	Alpha	Weighting system
0	0.35	Large domains
25	0.37	–
50	0.40	Medium domains
75	0.42	–
100	0.45	Specific domains
150	0.50	–
200	0.55	–
250	0.60	–

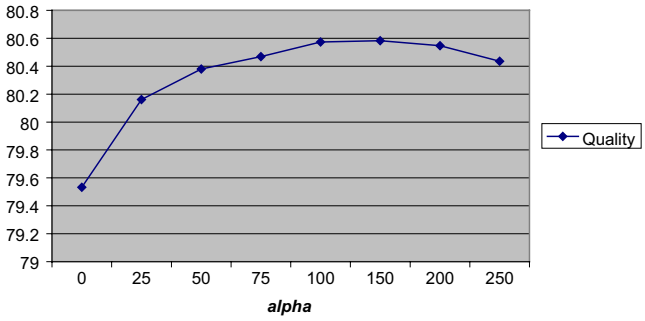

Fig. 1. First time slice (1991–1995): the quality criterion in relation to the alpha parameter (when the weighting system parameter is set to specific domains).

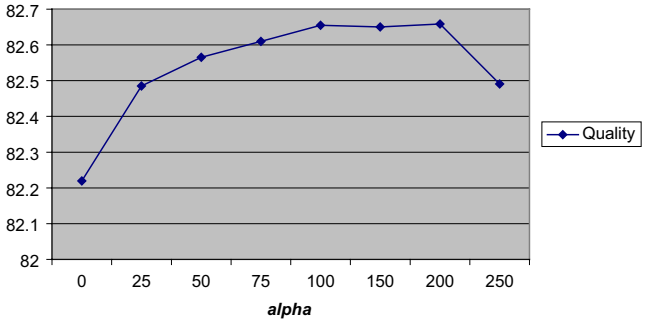

Fig. 2. Second time slice (1996–2000): the quality criterion in relation to the alpha parameter (when the weighting system parameter is set to specific domains).

Fig. 1 shows the quality criterion in relation to the alpha parameter for the first segment, while Fig. 2 shows the same for the second segment, in the case where the weighting system parameter is set to specific domains.

Both Figs. 1 and 2 seem to indicate the presence of a threshold value related to the alpha parameter. Exceeding this value, the quality criterion slowly starts to decrease, probably due to some kind of "overkill" effect.

On the grounds of these preliminary calibration studies, for our final clustering (and for both time slices) we decided to adopt the settings summarised in Table 5. In this respect, it is to be noted that, even if it made sense to allow the algorithm to enjoy more "freedom" during the calibration phase, for the final clustering sessions the maximum number of clusters allowed parameter was reduced to 20 in order to enhance the meaningfulness of the final patent maps, as our datasets were not very large.

Table 5
Setting of parameters for final clustering sessions

Maximum number of clusters allowed	Weighting system	KDT	Alpha
20	Specific domains	5	0.45

Table 6
Resulting values of built-in criteria

	Quality	Between	Within
First time slice (1991–1995)	80.99285	17.57098	65.38984
Second time slice (1996–2000)	84.01455	14.33116	62.90182

We also set a number of different stop words for both segments, which served to improve the overall clustering quality. The resulting values of the within, between and quality criteria for the final clustering sessions are shown in Table 6.

Regarding the internal built-in criteria, we concluded that, even if they do show some common behavioural pattern, their connection to a given dataset is rather strong, and that they should be considered more as generic quality indicators than exact means for validating a clustering process a priori.

3.3. Analysis of a patent portfolio: the validation step

The next logical step was to try to find a suitable methodology for validating the patent maps obtained with the PackMOLE™ prototype.

When a preexisting classification of the documents is already present, a number of metrics, for example the *entropy* [8], the *purity* [9], or the *gain ratio* [10], are well known in the field of document clustering for helping the analyst to perform the validation step.

Unfortunately, on a general perspective, the problem with these kinds of metrics is that they do not provide a measure of the intrinsic quality of a document clustering, rather they show the degree of alignment between the clustering and the preexisting classification.

When it comes to interpreting the results of a clustering, the effective usefulness of these metrics for competitive intelligence applications is therefore unclear. Moreover, to correctly assess, for example, the entropy of a clustering, it is necessary for each document to be associated with one class only, which is not usually the case with patents.

In order to overcome these issues, we had to choose our own validation criteria on the grounds of our understanding of the subject matter involved, and then check the consistency of every cluster in the bubble maps against these criteria.

Therefore it was decided that, for a cluster to be considered valid, at least 50% of its documents had to be found to be homogeneous, where the said homogeneity had to be intellectually assessed. We also decided to discard a cluster if it consisted of less than three documents, or if it presented wrong links with other clusters.

The valid clusters were further grouped into "regular" clusters and "borderline" clusters, the latter group consisting of clusters having a percentage of precisely 50% homogeneous documents.

The validation step was carried out by reading the documents in each cluster, as well as by relying, at least in part, on the available metainformation. The invalid clusters were then removed from the bubble maps. Only after the invalid clusters were removed, were we able to correctly extract the information about the patenting trends.

The final, validated (and graphically retouched) patent maps are shown in Figs. 3 and 4 for the first and second time slices, respectively.

The numbers appearing in each bubble correspond to the numbers of Derwent records contained in each cluster.

According to our criteria, 70% of all the clusters in the first segment were found to be valid, while in the second segment a slightly higher percentage of 75% valid clusters was observed. These results are summarised in Tables 7 and 8 for the first and second time slices, respectively.

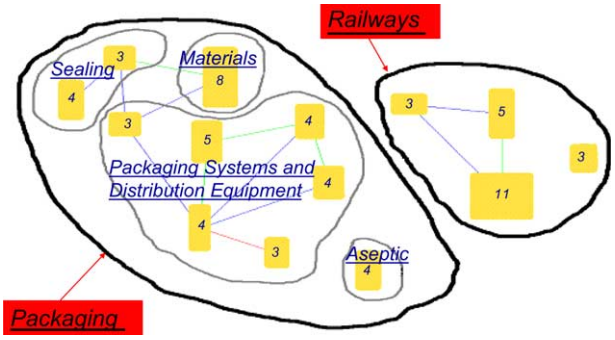

Fig. 3. First time slice (1991–1995): patent map.

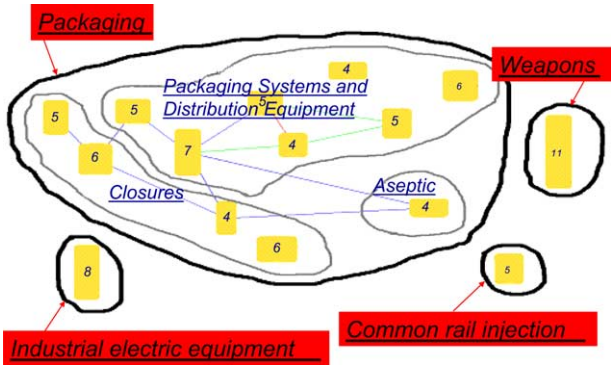

Fig. 4. Second time slice (1996–2000): patent map.

Table 7
Validation of clusters: first time slice (1991–1995)

Valid clusters		Rejected clusters
Regular clusters	Borderline clusters	
10	4	6

Table 8
Validation of clusters: second time slice (1996–2000)

Valid clusters		Rejected clusters
Regular clusters	Borderline clusters	
12	3	5

4. Discussion of results: comparison of clustering analysis and classification-based analysis

As mentioned above, the PackMOLE™ prototype is able to export data regarding the outcome of a clustering session to an external software program for further processing, namely a spreadsheet such as, for instance, Microsoft® Excel.

Using this feature, the first 10 most numerous Derwent Classes [11] were extracted from both time slices, as shown in Table 9, to study the patenting activities of our target company with a classical grouping technique based upon patent classifications, and then to compare the similarities and differences of this standard analysis with the patent maps previously obtained through the clustering sessions.

Clearly, even though the two different analyses more or less show the same overall trends, the patent maps obtained through text mining are easier to understand, in part because they are presented in graphical rather than textual or tabular form.

In fact, the patent maps generated by text clustering allow for a better overview of the relationships between the different areas of patent activity, at the same time avoiding the work involved in using different, more detailed patent classifications, such as for example the IPC. In particular, the IPC was felt to be either too broad (at the class/subclass level) or too detailed (at the group/subgroup level) to effectively carry out an optimal patent portfolio analysis.

Regarding the Derwent Classification, it is to be noted that the majority of retrieved Derwent Classes belonged to the Engineering sections P and Q where each Derwent Class automatically corresponds to an exact, predetermined range of IPCs [11].

Table 9
Classical analysis through patent classifications

Derwent Class	1991–1995	1996–2000	Description
Q31	36.0465	29.2453	Packaging, labelling
Q32	16.2791	22.6415	Containers
Q34	19.7674	15.0943	Packaging elements, types
Q35	15.1163	14.1509	Refuse collections,conveyors
A92	9.30233	12.2642	Packaging and containers
Q79	0	12.2642	Weapons, ammunition, blasting
X25	9.30233	8.49057	Industrial electric equipment
P72	9.30233	7.54717	Working paper
Q39	0	7.54717	Liquid, handling
P62	6.97674	5.66038	Hand tools, cutting
Q21	18.6047	0	Railways
Q11	0	6.97674	Wheels, tyres, connections

Note. Numeric values represent percentages. As one would expect, the total percentages are greater than 100, as patents are usually labelled with more than just one Derwent Class.

This subdivision scheme did not always prove effective: for instance, in some extreme cases a few patents were classified with different IPCs, even if they clearly referred to inventions sharing the same subject matter, [5] and their respective IPCs were spaced so far apart from each other that they were assigned different Derwent Classes as well. On the contrary, the PackMOLE™ prototype was able to correctly group these patents into the same clusters.

In any event, the good performances exhibited by the PackMOLE™ prototype in correctly grouping patent documents were probably greatly enhanced by the high quality of Derwent abstracting.

Another strong point shown by the PackMOLE™ prototype was to provide the analyst with the ability to correctly identify, by comparing the two patent maps shown in Figs. 3 and 4, the changes in patent activities in the different business areas of our target company, as well as the subtle dynamics related to technological developments and spin-offs, that were not otherwise immediately detectable through the classical analysis: Indeed, by only relying on the results shown in Table 9, one might have deduced quite opposite conclusions.

Finally, it is worth noting that the validated patent maps shown in Figs. 3 and 4 do not exactly represent the complete patent portfolio of our target company, as a few clusters were deemed to be invalid and thus removed, as previously mentioned.

Rather, the information contained in the validated patent maps was thought to constitute a fairly good picture of the patenting trends of our target company.

5. Final considerations

The strengths and weaknesses of the PackMOLE™ prototype were evaluated.

The tool showed a series of interesting advantages over classical patent portfolio analysis techniques, and proved to be effective in a real business scenario.

In many respects, text mining technology lets the analyst overcome the limits of current patent classifications.

On the other hand, the calibration and validation steps of the clustering process itself proved to be difficult, time-consuming and strongly dependent upon the contents of each dataset.

Probably, text mining techniques and patent classifications should not be considered alternative tools for patent mapping: Rather, they should be used in synergy.

Patent classifications, for instance, certainly have the potential to help the user during the validation step of a clustering session more than any built-in criteria.

Therefore, the next generation of text mining tools for patent analysis should integrate some kind of facility for manipulating patent classifications and other descriptive indexes or terms, to speed up the whole process, at the same time guaranteeing a professional quality of results.

Acknowledgements

Tetra Pak Carton Ambient SpA and CINECA are founding members of the CRIT (Centro di Ricerca e Innovazione Tecnologica) [6] consortium. Their cooperation in researching business-grade text mining applications for patent analysis was therefore conducted under the auspices of CRIT.

[5] Apart from the obvious issue regarding the different editions of the IPC, there is also the well-known problem of different patent offices applying the IPC in different and not always consistent ways.

[6] CRIT is located in viale Mazzini 5/3, 41058 Vignola (Modena), Italy.

References

[1] Cabena P, Hadjinian P, Stadler R, Verhess J, Zanasi A. Discovering data mining: from concept to implementation. Englewood Cliffs, NJ: Prentice Hall; 1997.
[2] Hehenberger M, Coupet P. Text mining applied to patent analysis. Paper presented at the 1998 Annual Meeting of American Intellectual Property Law Association (AIPLA), October 15–17, Arlington, VA.
[3] Krier M, Zaccà F. Automatic categorisation applications at the European patent office. World Patent Inf 2002;24(3):187–96.
[4] Smith H. Automation of patent classification. World Patent Inf 2002;24(4):269–71.
[5] Hull D, Aït-Mokhtar S, Chuat M, Eisele A, Gaussier E, Grefenstette G, et al. Language technologies and patent search and classification. World Patent Inf 2001;23:265–8.
[6] Trippe A. A comparison of ideologies: intellectually assigned co-coding clustering vs ThemeScape automatic thememe mapping. In: Proceedings of the 2001 Chemical Information Conference.
[7] Grabmeier J, Rudolph A. Techniques of cluster algorithms in data mining. Version 2.0. IBM Informationssysteme GmbH; 1998.
[8] Shannon C. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423, and 623–56.
[9] Zhao Y, Karypis G. Criterion functions for document clustering: experiments and analysis. Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001.
[10] Quinlan R. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.
[11] Derwent Information, Derwent World Patents Index®, the Derwent Classification, Edition 2. May 2000.

Michele Fattori obtained a 5 year degree (M.Sc. equivalent) in Materials Engineering from the University of Modena. Prior to joining the Intellectual Assets Department of Tetra Pak Carton Ambient in 2001, he worked as trainee patent and trade mark attorney for a leading Italian consultancy firm.