Decision tree analysis as a tool to optimise patent current awareness bulletins

Elif Doğanşargil¹, Michele Fattori *

Basell Poliolefine Italia S.r.l., P.le Privato G. Donegani 12, 44100 Ferrara, Italy

Abstract

Efficient setting up and maintenance of patent current awareness bulletins requires a good and constant dialogue between the information professionals and the end users. In view of giving the end users a chance to participate in the definition and optimisation of the patent bulletins, at the same time trying to simplify and rationalise as much as possible such a time consuming and labour-intensive task, the use of decision tree analysis was investigated, as a promising solution for determining a classification model capable of helping the information professional in assessing the desired level of detail to be assigned to each specific search profile.

Keywords: Patent bulletins; Current awareness; Dissemination of patent information; Decision tree analysis

1. Introduction

Current awareness bulletins are well-known in the patent information industry. These bulletins are widely used as a means to disseminate patent information among a multiform audience of end-users, such as scientists, researchers, engineers, managers, and so on.

Many industrial corporations, and Basell Polyolefins² and its predecessor companies are no exception, have a long tradition of preparing such bulletins in-house for subsequent dissemination among interested colleagues. Gone are the days when these bulletins were printed on paper, then bound and circulated to colleagues, before being stored in a library shelf: today, patent awareness bulletins are conveniently being distributed either as an email attachment to selected groups of end users or through the corporate intranet.

One of the main problems of patent current awareness bulletins resides in the fact that they are normally put together, as their name implies, as bulletins intended for alerting the readers of new patent publications in the name of competitors or related to specific technical areas. Once read and digested, the bulletins are put away and forgotten. The use of such bulletins as data feeds for building customized, in-house databases that can be efficiently searched at a later time, even though in theory possible and certainly in many cases also advisable, does not normally constitute their primary raison d’être. This “paper age” legacy limitation seems more a matter of perception than of substance: in most cases, currently available software solutions already allow patent bulletins to be easily collected, indexed and stored so as to enable efficient retrieval and future reuse.

Another aspect which is certainly typical of patent current awareness bulletins lies in their innate ability of being crafted to become highly tailored to the end users’ needs. While this is an obvious benefit, at the same time it should be taken into account that such an approach implies the presence of a labour-intensive process that, if not properly structured and closely monitored, may easily become too complex, cumbersome and, eventually, substantially unmanageable [1].

---

¹ Corresponding author. Tel.: +39 0532 46 7300; fax: +39 0532 46 7675.
E-mail addresses: Elif.doganavsargil@basell.com (E. Doğanşargil), Michele.fattori@basell.com (M. Fattori).
The present article explores the use of decision tree analysis as a means to correctly define a classification model capable of helping patent information professionals in assessing the desired level of detail, namely the granularity and degree of customization of each search profile included in a multi-profile patent awareness bulletin, in a structured and flexible way.

This work was originally carried out as part of a dissertation thesis developed by one of the authors in the course of the Master in Science, Technology and Management (MaSTeM Course) during the academic year 2006–2007.

MaSTeM courses are the fruit of a long standing and successful cooperation between the University of Ferrara and Basell Polyolefins.

2. Preliminary internal survey

2.1. The patent current awareness bulletins

Within Basell, patent current awareness bulletins, issued on a monthly basis, are prepared and distributed by the Documentation Services department which is an integral part of the Intellectual Property function.

The bulletins, which physically consist of an MS Word file containing title, abstract and bibliographic details of new patent publications, are classified within nine different search profiles, each devoted to a specific subject matter, ranging from basic research in the field of catalysts to commercial applications of polymeric materials in various technical fields.

Only one patent database is currently used to create the bulletins, namely the Derwent World Patents Index, provided by Thomson Scientific. Each search profile comprises a selection of both keywords and patent classifications in order to fully harness the flexibility of the many indexing levels provided by the database, from the latest revision of the International Patent Classification to Derwent’s proprietary Polymer Indexing [2].

Generally speaking, the search profiles are built to retrieve as many results as possible – thus putting an emphasis on recall rather than precision. The results are then manually checked and pruned by the information analysts and only those patents that are deemed to be relevant are eventually collected in the bulletins. Depending on the kind of search profile, the percentage of relevant patents may range from 30% to 70% out of the total set of retrieved documents, with the profiles dealing with applications being those that include the most background noise.

At the end of the manual checking process, each profile may contain on average from 5 to 60 patents. Understandably, this recurrent manual checking is quite an onerous task.

2.2. The survey: rationale and results

In view of obtaining a clear understanding of the needs and informational behaviour of the end users, namely of the internal customers of the patent current awareness bulletins, it was decided to carry out a preliminary survey [3,4].

Such a survey was felt to constitute a necessary step in order to measure the effective usage of the patent bulletins, to seek alignment between the efforts of the Documentation Services team and the needs of the end users, and also to obtain some practical information that could be used to optimise the different patent search profiles that make up the bulletins.

In January 2007, Documentation Services prepared a multiple-choice questionnaire, which was designed to be as simple and unambiguous as possible.

The survey was directed at acquiring sufficient insight on many important facets of the bulletins, such as format, content, frequency, usage, accessibility and so on. Additionally, free-text comments and feedback were also encouraged. Broadly speaking, we aimed at gauging the users’ interest in the bulletins (both in general and more specifically, with reference to particular sections of the bulletins).

The survey involved all Basell R&D research staff from seven target work groups, namely 242 people. In this article, we represent work groups by the letters A–G.

Each work group is devoted to conduct R&D activities in a different, specific area of polyolefin technology. The questionnaire was distributed via email to researchers. The email message containing the questionnaire was followed up by some reminders.

At the scheduled deadline, 147 answers were received, which represent a total response rate of slightly over 61%. Table 1 presents the distribution of the actual survey participants expressed in terms of frequency of response and percentage for each work group.

In order to carry out simple descriptive statistical analysis and to summarize the structure and distribution of our data, we used the MINITAB6 software package. Furthermore, as the data obtained from the survey were qualitative, we created a coding form in MS Access which gave

---

Table 1

<table>
<thead>
<tr>
<th>Work group</th>
<th>Total (n = 147)</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>46</td>
<td>79</td>
</tr>
<tr>
<td>B</td>
<td>29</td>
<td>50.8</td>
</tr>
<tr>
<td>C</td>
<td>22</td>
<td>61.1</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>34.6</td>
</tr>
<tr>
<td>E</td>
<td>15</td>
<td>55.5</td>
</tr>
<tr>
<td>F</td>
<td>21</td>
<td>67.7</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>71.42</td>
</tr>
</tbody>
</table>

---

5 Doganavsargil.

6 MINITAB is a well-known computer software designed to perform basic and advanced statistical functions (http://www.minitab.com).
us the possibility to convert textual data into numeric data before inputting in MINITAB.

A detailed discussion of each and every aspect of the survey and its results is beyond the scope of this article. Instead, we will focus only on those aspects which are more closely related to the implementation of the decision tree analysis.

In particular, as mentioned earlier, the bulletins consist of nine different search profiles, which are identified in this article with Roman numbers I–IX.

Hence, we decided to investigate the preferences of each work group A–G with respect to each search profile I–IX contained in the bulletins. For this purpose, we asked them which profile (or profiles) they preferably read.

A correlation between work groups and their preferred search profiles is shown in Table 2 expressed in terms of percentage. A “preference rate” was accepted as significant if it received at least 30% of the expressed preferences; such significant values are bolded in Table 2.

According to survey results and with reference to Table 2, search profiles III and IV appear to be the most popular ones, both ranking high in the preferences of more than three work groups at the same time. In particular, profile IV appears to be the most preferred search profile of all. Search profiles I, II, V and VI also fare reasonably well, whereas the remaining profiles seem to be directed at satisfying information needs in more specific areas.

We also note that colleagues belonging to work group A seem to be the most demanding ones, showing considerable interest in many different profiles. Finally, we note that work groups C, F and G showed a strong preference (with rates varying between 40% and 86%) for profiles I–III.

Another important aspect of the survey was to collect feedback regarding the degree of satisfaction of the users with what can be called, broadly speaking, the “level of detail”, or granularity, of the search profiles. Such level of detail is of course directly linked with recall and precision, both parameters being ultimately determined, for each search profile, by the information analyst, on the basis of his/her understanding of the users’ needs. In other words, the balance between quantity and quality of the information included in the bulletins must be appropriately set by the analyst, also taking into account related factors such as content, format and size.

Among all the respondents, 27.9% of the users interviewed were satisfied with the search profiles as they are currently structured. At the same time, 32% of respondents would prefer to receive “more generic” search profiles, whereas 24.5% would opt for “more specific” profiles.

These findings seem to suggest that, in terms of satisfaction about the level of detail of the search profiles, the users can be segmented in three, almost equally important, groups: the first group comprises people who are satisfied with the current search profiles, whereas the second and the third groups comprise people who would prefer more generic and more specific profiles, respectively.

2.3. Important implications of survey results

As a consequence of the above results, it appears that having identified, for each work group, which are the most popular search profiles, is not enough.

It is also decidedly important to understand, for each profile and for each work group, which is the desired level of detail, to enable the information analysts to adequately set a proportion between precision and recall which is consistent with the users’ needs.

Before dealing with this fundamental issue, it is possible to draw a few other useful pieces of information from the survey results.

In fact, the survey included questions intended to investigate whether the bulletins are read, with which frequency, and if not, for which reasons. About two-thirds of respondents said that they read the bulletins, with varied frequency. However, among those two-thirds of readers, those who qualified themselves as occasional users also said that they would like to read the bulletins more often if only they had enough time.

It’s interesting to combine this finding with the outcome of the free-text feedback form\(^7\) included at the end of the questionnaire. This was the only part of the survey that

---

\(^7\) In most cases, this was followed up by telephone interviews.
was not structured as a multiple-choice questionnaire, thus
giving the users a chance to freely express their suggestions.

Generally speaking, four main groups of suggestions
emerged, and namely: the need for a more understandable
format and easier access to the bulletins, the addition of
new search profiles, the inclusion of a monthly comment
to highlight what is really interesting and new, and finally
the integration of non-patent literature as well.

Rather obviously, a better and more understandable
format could certainly enable users to access the bulletins
in a more efficient and less time consuming way. The need
for new search profiles could also help raise interest in the
patent bulletins among those who are not regular readers
or do not even currently read the bulletins at all. The two
last comments seem more related to the diversification of
needs expressed by users in areas that tend to be closer to
the business environment or to scientific research,
respectively.

Furthermore, it is to be noted that also those who
defined themselves as regular readers expressed the need
for an easier to access and more efficient format.

Clearly, the format of the bulletins and the new search
profiles must be crafted taking into account the desired lev-
els of detail as expressed through the survey.

According to definitions well-established in the relevant
literature [5], it appears that the need for “more generic”
search profiles refers to the need of having what can be
defined as “macro-level” patent analysis, whereas “cur-
rent” or “actual” profiles refer to “meso-level” patent anal-
ysis and, finally, “more specific” profiles refer to “micro-
level” patent analysis.

Therefore, as exemplified in Fig. 1, “macro-level” anal-
ysis will primarily deal with large numbers of documents,
“micro-level” analysis with limited numbers of documents,
with “meso-level” analysis somewhere in between.

It is the opinion of the authors that micro- and meso-
level analyses can be dealt with, in most cases, with sub-
stantially the same tools, namely the usual range of online
services offered by various information providers, where
macro-analysis probably needs the intervention of some
kind of data or text mining approach [6–8].

In any event, the real challenge for patent information
departments is that in some cases the users will need multi-
ple versions of the same bulletins, depending on the desired
level of detail, hence increasing the workload brought to
information analysts.

In the following sections of this article, a technique
known as decision tree analysis will be applied to solve this
issue.

3. Decision tree analysis

3.1. Why decision trees?

Decision tree analysis is a formal, structured approach
which eases the knowledge-acquisition for decision mak-
ing. Decision trees help decompose a complex problem into
smaller, more manageable undertakings which allow the
decision makers, in our specific case the information ana-
lysts, to make smaller determinations along the way to
achieve optimal overall decisions [9].

The attractiveness of decision trees is that they acquire
‘classification knowledge’ that might be used for pattern
classification. Such knowledge, put in the form of a deci-
sion tree or a set of rules, can be used to estimate answers
or class variables for a given new case. Decision trees enjoy
several advantages. As they are a tool which uses rules,
they help decision makers make decisions in fields where
they are faced with difficulties in formalizing knowledge.
Moreover, when used to handle a given classification case,
decision trees not only provide the solution for that case,
but also clearly state the reasons for each choice that is
made [9].

Decision trees can handle big amounts of data. Their
representation of acquired knowledge in tree form is intu-
itive and generally easy to assimilate by humans [10].

The structure of a decision tree is similar to that of a real
tree with a root which represents the decision problem,
with each branch representing a classification rule, and
with so-called chance nodes and decision nodes indicating
the value of target classes and the outcome states,
respectively.

Constructing classifiers from data in the form of deci-
sion trees has been quite popular and successful real-world
applications in different fields that employ decision tree
analysis have been reported. For instance, in banking they are used to make credit decisions, in industry for the diagnosis of mechanical devices, e.g. in production quality control, in healthcare for diagnostics of some diseases, in molecular biology to analyze amino acids composition and in astronomy to automatically classify celestial objects. These are only some of successful real world applications [11]. It is worth noting that in some cases, decision trees have already been applied to patents and patent searching, albeit with different objectives and scope [15,16].

It is obvious that many other potential applications have not yet been explored.

### 3.2. Determining the appropriate level of detail

A decision tree algorithm is used to construct a decision tree classifier for determining an appropriate class (among a predetermined set of classes) for a given case to be tested [9,12].

The goal of this classification task is to discover meaningful relationships in a body of training data presented as individual examples (usually called instances), and to produce a generalization of those relationships that can be used to interpret subsequently presented data.

In the training data, each instance (often presented as a row of data in a table) comprises a set of attributes and is put into relationship with one or more classes (classes are presented as column entries for each instance) [12].

In our case, the attributes were, of course, the preferred levels of detail as expressed by the interviewed users, for each search profile. The available classes were the following: GENERIC, ACTUAL and SPECIFIC, corresponding to more generic, current or more specific search profiles, in other words to macro-, meso- and micro-analysis, respectively.

One standard approach for the induction of rules involves dividing the data into two sets, then performing training on the first set, and finally testing the induced knowledge on the second set [13,14]. One can repeat this process a number of times with different splits, then calculate the average of the obtained results to estimate the rules' performance on possible new test data.

In the case of determining the appropriate level (or levels) of detail to be assigned to each search profile, an excerpt of a training dataset is given in Table 3.

In the training process, the first step is to construct a classification model as a set of rules induced from training examples. Rule induction from examples is a machine learning technique that has been successfully used as a support tool for knowledge acquisition and prediction. Classification rules will then be expressed as a decision tree which is constructed beginning from the root of the tree and going through the leaves. An important point when constructing classification rules is that such rules classify any unseen possible future example as well as the examples of the training set. Hence the decision tree must reflect meaningful relationships between each class of each instance and the values of its attributes [13].

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Class, level of detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual profiles (%)</td>
<td>Specific profiles (%)</td>
</tr>
<tr>
<td>33</td>
<td>11</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>9</td>
</tr>
</tbody>
</table>

Rules are made up of two major parts: a first part that comes between the operators ‘IF’ and ‘THEN’, expressing a condition, and a second part which comes after the operator ‘THEN’ defining a class that, in our case, corresponds to one or more levels of detail.

The following rules 1–4 were generated according to the experienced relationships of levels of detail in terms of type of analysis to be dealt with, through examples extracted from the survey that were treated as training data. As can be seen, conditions and classes define logical expressions:

**Rule 1:** IF \[\sum (\% \text{ Actual profiles} + \% \text{ Specific profiles}) \geq 50\% \text{ AND } \% \text{ Generic profiles} < 40\%\] THEN Check for dominancy between \% Specific profiles and \% Actual profiles

IF \% Specific profiles \geq \% Actual profiles

THEN Profile = SPECIFIC

IF \% Specific profiles < \% Actual profiles

THEN Profile = ACTUAL

**Rule 2:** IF \[\sum (\% \text{ Actual profiles} + \% \text{ Specific profiles}) < 50\% \text{ AND } \% \text{ Generic profiles} \geq 40\%\] THEN Profile = GENERIC

**Rule 3:** IF \[\sum (\% \text{ Actual profiles} + \% \text{ Specific profiles}) \geq 50\% \text{ AND } \% \text{ Generic profiles} \geq 40\%\] THEN Check for dominancy between \% Specific profiles and \% Actual profiles

IF \% Specific profiles \geq \% Actual profiles

THEN Profile = SPECIFIC and GENERIC

IF \% Specific profiles < \% Actual profiles

THEN Profile = ACTUAL and GENERIC

**Rule 4:** IF \[\sum (\% \text{ Actual profiles} + \% \text{ Specific profiles}) < 50\% \text{ AND } \% \text{ Generic profiles} < 40\%\] THEN Check for dominancy between \% Specific profiles, \% Actual profiles and \% Generic profiles

IF Specific profile is dominant

THEN Profile = SPECIFIC

IF Actual profile is dominant

THEN Profile = ACTUAL

IF Generic profile is dominant

THEN Profile = GENERIC

As the levels of detail of the search profiles relate to the type of analysis to be carried out (to be chosen between macro-, meso- and micro-analysis), we generated an additional rule that defines the determined class in terms of...
quantity of patent documents. This is rendered necessary because, in the case of GENERIC profiles, a macro-analysis can not be carried out in a sensible manner if the number of available documents is too low. The additional rule does not make part of the main structure of the tree but helps in interpreting the correct action to be performed within each class:

**Additional Rule:**

IF Profile = SPECIFIC

THEN Supply and publish monthly title and abstract

IF Profile = ACTUAL

THEN Supply and publish monthly title and abstract

IF Profile = GENERIC AND average number of patent applications in a month ≥ 50

THEN Supply and publish monthly statistical analysis of patent information

IF Profile = GENERIC AND average number of patent applications in a month < 50

THEN Supply and publish triannual and/or semianual statistical analysis of patent information.

As can be seen, we fixed a threshold level of 50 documents, a value below which a macro-analysis is deemed not to make much sense. Of course, this threshold level is only provided as a mere indication and can be modified depending on needs. It is also possible to define multiple threshold levels, in order to help in deciding which statistical, graphical, data and/or text mining tools are to be used, depending on the quantity of documents that should be analysed.
A graphical visualization of the decision tree built upon rules 1–4 plus the additional rule is given in Fig. 2.

A rule set can be easily converted into a decision tree by placing one rule for each path from the root to a leaf and vice versa.

Let’s discuss the testing process with a specific example of application of the rules: in the case of search profile IV (see Table 4), the rules would classify the first instance (namely, the instance related to work group A), as GENERIC by Rule 2, the second (related to work group B) as both ACTUAL and GENERIC by Rule 3, the third (related to work group D) as GENERIC by Rule 4, the fourth (related to work group E) as SPECIFIC by Rule 1, the fifth (related to work group F) as both SPECIFIC and GENERIC by Rule 3 and the last instance (related to work group G) as GENERIC by Rule 2.

Table 5 shows the overall results obtained by applying the decision tree model to all the data set, namely to all the search profiles I–IX, which were deemed to be interesting enough by the work groups (as per results shown in Table 2), for each work group A–G.

4. Discussion

According to survey results, one of the main reasons that discourage both current and potential users of patent information in reading the current awareness bulletins lies in the fact that the search profiles do not focus well on their changing information needs. The users clearly expressed the need for new search profiles, as well as the need of being provided with patent information at different levels of detail. These findings brought us to the decision of assigning different levels of detail to each specific search profile. In fact, in order to appropriately suit the users’ needs, a few profiles received more than just one level of detail.

Decision tree analysis was investigated as a promising tool to objectively determine, in a structured way, which level of detail has to be assigned to each profile.

Other problems were highlighted by the survey, especially regarding the format of the bulletins which seems to further discourage the use of patent information. A significant number of respondents mentioned their difficulties in accessing patent information and in understanding the format of the bulletins.

The reformulation of the search profiles as indicated by the outcome of the decision tree analysis, the addition of new profiles, and the adoption of more modern, efficient software platforms for the collection, analysis, review and dissemination of the bulletins, may certainly help in solving the issues highlighted by the survey.

The results that were obtained are thus deemed to be satisfactory, in the sense that a simple, understandable classification model has been codified, which model let the analysts formalize knowledge and hence help them base their decisions on the grounds of a formal structure.

Various applications of decision tree analysis, in different fields, are reported in the scientific literature. It is to be noted that, most of the time, these applications deal with large datasets [3].

In our case, however, the datasets contain a very small number of examples due to the nature of the classification problem. It is obvious that, in the future, depending on the ever changing needs of the end users, the search profiles and their corresponding levels of detail may be adjusted accordingly. In such a case, owing to the availability of more data, one can also attempt to modify and improve the decision tree in the light of new examples.

5. Conclusion

Clearly, one of the principal strengths of patent current awareness bulletins is that they can be easily customised to meet the end users’ needs.

The idea behind this study was to try and define some degree of structure in the amount of customisation that
must be implemented in the bulletins, also trying to minimise the impact on the workload that such a customisation inevitably brings to the information analyst.

One of the main problems the information professional is likely to encounter during the customisation process lies in the fact that different users of the same search profile may have different needs in terms of desired level of detail. To cope with this apparent contradiction, it is important to be able to customise such level of detail or granularity of the search profiles, thus providing different users with different versions of the patent awareness bulletins, according to a defined set of rules which in our case were successfully described through application of decision tree analysis.

It is also important to understand that not only the content, but also the format of the bulletins must be adjusted depending on the desired level of detail, and this consideration implies that to prepare patent current awareness bulletins in a modern, customised way, the information analyst must be flexible enough to consider using different tools and techniques.

References


Elif Doğanavşargil obtained a degree in Industrial Engineering from Dokuz Eylül University (Izmir-Turkey) in 2000 and a Master degree in Science, Technology and Management (MaSTeM) in 2007 from the University of Ferrara, with a thesis titled ‘Best practices in patent information retrieval, selection and dissemination in a multinational company’, which was awarded with the XVI FOSCHINI supplementary prize. Prior to that, she has worked as production engineer in Turkey, specializing in the planning and optimisation of production processes. She is currently working at Basell Polyolefins, dealing with a project focused on innovating the distribution of patent information.

Michele Fattori currently heads the Documentation Services department of Basell Polyolefins. He is the founder and past president (2004–2007) of the AIDB, the Italian Patent Information Users Group and presently serves as vice-chair of the Confederation of European Patent Information Users Groups. Prior to joining Basell, he held positions related to patent information and intellectual property in both industry and private practice. Michele obtained his M.Sc. in Materials Engineering from the University of Modena and Reggio Emilia.