Analysis of Patents for Prior Art Candidate Search

Sébastien Déjean*, Nicolas Faessel†, Lucille Marty‡, Josiane Mothe†, Samantha Sadala‡ and Soda Thiam‡
*IMT, Toulouse, France
Email: sebastien.dejean@math.univ-toulouse.fr
†IRIT, Toulouse, France
Email: {faessel,mothe}@irit.fr
‡INSA, Toulouse, France
Email: {lmarthy,sadala,thiam}@etud.insa-toulouse.fr

***Abstract*—In this paper, we describe a method for analyzing a collection of patents in order to help prior art candidate search in an interactive and graphical way. The method relies on the use of two data mining methods: hierarchical agglomerative clustering and principal component analysis, which are applied successively. The correlation between the application patent and the other patents is a good indicator to help decide the classes of patents to look at.**

***Index Terms*—Information retrieval, Patent retrieval, Prior art retrieval, Visualization, Information mining.**

I. Introduction

Patents correspond to one type of intellectual property right that plays an important role in innovation and in the economy. Patent retrieval is crucial considering the amount of existing information and economic issues associated with patents. The number of patents available make mandatory to have effective and efficient ways of searching and browsing them. For example, Thomson Scientific provides the Derwent World Patents Index® , which ”contains over 21.85 million patent families covering more than 45.2 million patent documents, with coverage from over 47 worldwide patent authorities” [1]. Additionally, when applying for a new patent, prior art should be verified both by the applicant, to fill in the application, and by the certifier to analyze the application. 1.98 million applications were filed in worldwide in 2011 according to the World Intellectual Property Organization (WIPO) [2]. Verifying prior art is thus crucial and has been investigated in the relevant literature. International evaluation forums also define tasks to evaluate prior art search.

The Conference and Labs of the Evaluation Forum Initiative (CLEF), formerly known as Cross-Language Evaluation Forum, launched the Intellectual Property track (CLEF-IP) in 2009 to investigate Information Retrieval (IR) techniques for patent retrieval. In 2011, CLEF-IP specifically focuses on prior art candidate search. The Text REtrieval Conference Chemical Track 2011 (TREC-CHEM 2011) also considers prior art candidate search [3]. Prior art search is not the only task evaluation forums investigates, there are many information retrieval and analysis tasks and other challenges related to patents [4], [5]. However, in this paper, we will focus on this task.

Prior art candidate search aims at querying and retrieving the patents in order to discover any knowledge existing prior to the analyzed patent application [6]. The underlying objective that is pursued is to find patents, which can invalidate a given patent application. Most of the approaches to this task use standard information retrieval (IR) processes [7], [8].

Results are evaluated using standard IR measures that consider the rate of relevant documents that have been retrieved and the rate of retrieved documents that are relevant. A relevant document in the case of prior art candidate search is a patent that is evaluated as potentially invalidating the current patent application. More specifically, evaluation measures such as Mean Average Precision (MAP), recall, precision, recall at 100 (when 100 retrieved documents are considered) and precision at 100 have been used.

In this paper, we make post evaluation analysis. More specifically, this study aims at analyzing patents knowing the relevance judgments. We show that considering patent title and patent abstract is complementary. We also show that it is possible to iteratively reduce the number of patents to be analyzed while keeping a high level of recall. Patent analysis is done using clustering and factorial analysis methods. The analysis results in a graphical visualization the user can interact with. The rest of this paper is organized as follows. Section II presents related works regarding prior art search, and browsing and clustering document collections. Section III describes the way patents are indexed and the resulting representations to be analyzed. Section IV presents the methodology of analysis and Section V the results of the patent analysis. Finally, Section VI discusses the results and concludes this paper.

II. Related Work

A. Prior art candidate search

Prior art candidate search has been studied in CLEF-IP as finding ”patent documents that are likely to constitute prior art to a given patent application” [9]. Figures 1 and 2 show an application patent, and a previous patent potentially invalidating the application.
The results obtained in the evaluation campaign show that the task is quite hard. For example, the best run in 2010 obtained a MAP of about 0.26 [7].

Mainly, standard IR methods have been used in the literature to solve this task. Madgy et al. [10] have used standard information retrieval techniques (stop word removal, stemming

ISBN: 978-1-61208-311-7

```
<patent−document ucid=”EP−1236420−A1” lang=”FR”>
  <bibliographic−data>
    <technical−data status=”new”>
      <invention−title lang=”DE”>Brste zum Auftragen eines
           Produktes auf keratinische Fasern
       </invention−title>
      <invention−title lang=”EN”>Brush for applying a
           product on keratinous fibres</invention−title>
      <invention−title lang=”FR”>Brosse pour l’application
           d’un produit sur les fibres kratiniques
       </invention−title>
    </technical−data>
  </bibliographic−data>
  <abstract lang=”EN”>
    <p>The applicator comprises a rod with a brush at one
         end. The portion of the rod adjacent to the brush
         has an axis (Y) and the brush comprises a core (11)
          from a portion of which bristles extend. The core
         is curved over a part of its length and the angle
         between the rod axis and the core axis is less than
          90o and the brush free end is not aligned with the
          rod axis.</p>
  </abstract>
  <abstract lang=”FR”>
    <p>La prsente invention concerne un dispositif pour l’
         application d’un produit sur les fibres kratiniques
         , notamment pour l’application de mascara sur les
         cils, comportant une tige munie  une extrmit d’une
         brosse, la portion de la tige adjacente  la brosse
         ayant un axe (Y).</p>
  </abstract>
  <abstract lang=”FR”>...</abstract>
  <description lang=”FR”>...</description>
  <claims>...</claims>
</patent−document>
```

Fig. 1. Patent application

using Porters stemmer) and obtained a Mean Average Precision of 0.1216, a recall at 100 of 0.3036 and precision at 100 of about 0.228. They used the Indri system that ranks the results using a language model and inferred networks. Becks et al. [11] used the Okapi system and BM25 weighting. Most of the other participants also used either the Lemur/Indri system or the Apache/Lucene system. However, these approaches do not allow browsing and interactive visualization.

B. Browsing and clustering document sets

One frequently cited work regarding document browsing is Scatter/Gather [12]. The principle developed in this approach is an interactive refinement of the target document set. From an initial clustering of a document collection in k-clustering, the users select the clusters they are interested in; then, the system re-clusters this subset of documents, and so on.

Many works have investigated search result clustering either to re-rank results initially retrieved by a search engine or to group the results and provide users with clusters of documents they can choose [13].

Classification in the context of patents is generally concerned with classifying patents using the International Patent Classication (IPC) codes or other classification schema. However, some interesting work has been conducted in this context for patent mapping. Kim et al. [14] for example cluster patent document contents using k-means. From the clustering results, they extract a semantic network that helps having an overview of the patent subset. Sharma [15] uses k-means clustering algorithms in order to structure a patent sub-collection. Jun et al. [16] uses patent clustering in order to predict technology trends. Djean and Mothe [17] also presents various visual clustering methods and applications.

```
<patent−document ucid=”EP−0792603−A1” lang=”FR”>
  <bibliographic−data>
    <technical−data status=”new”>
      <invention−title lang=”DE”>Brste zum Anbringen von
           Kosmetika und insbesondere Mascara</invention−
           title>
      <invention−title lang=”EN”>Brush for applying
           cosmetics and in particular mascara</invention−
           title>
      <invention−title lang=”FR”>Brosse progressive pour
           appliquer un produit cosmtique, notamment du
           mascara</invention−title>
    </technical−data>
  </bibliographic−data>
  <abstract lang=”EN”><p>The applicator brush consists of a
        cylinder of bristles radiating from a core in the
       form of a twisted metal wire spiral, and has at least
        one concave curved recess (107) in the cylindrical
       surface. The recess is oval, circular or elliptical
       in shape, or it can be made from two sections of a
       circle which intersect. It is located in a zone (109)
        of the bristle cylinder which is less dense than the
        two ends (112,113). The ends of the cylinder have
       alternating long and short bristles. The applicator
       brush can be attached to the inside of a cap which
       screws onto the neck of a cosmetic product container
       .</p>
  </abstract>
  <description lang=”FR”>...</description>
  <claims>...</claims>
</patent−document>
```

Fig. 2. Prior art candidate to invalidate patent in Figure 1

Our work also uses document clustering and interactive refinement of clusters. More specifically, we cluster patents according to their content, which is automatically extracted. In addition, our work studies the effects of the clustering parameters on the results.

III. Patent representation for text analysis

Since the collection is composed of more than three million patents, we decided to first select a sub-part of these patents, the ones that are more likely to be prior art candidates for each topic. To select this subset of patents, we use a standard information retrieval process. Then, we built a representation that fits the type of analysis we intend to investigate. We generate a patent representation that keeps both the section part from which each term occurs and the language in which it is used. The collection and these two steps are described in this section.

A. Collection

The collection we used is the one used in CLEF-IP forum 2011 [18]. It consists of more than three million patent documents from European Patent Office sources with contents in French, German and English. Several documents can be related to the same patent and correspond to versions (e.g. application phase and granted patent). The documents contain bibliographic data, a title, an abstract, a description, and claims. The patent title is provided in three languages (English, German and French). For some of the patents, the abstract is also provided in the three languages, for others in one of the languages only. The description and claims are in one language only.

Topics correspond to patents and thus have the same structure. In this study, we used the 1000 official topics, to which are associated the patents relevant to the task. Relevance judgements are produced automatically, using patent citations from seed patents.

B. First patent selection

To make a first targeted sub-collection for each topic, we used the Terrier system [19]. We built an index considering the English titles and English abstracts. We used a stop-word list and Porter's stemmer and the default Terrier parameters. Then, from a topic, we query the index and retrieve, at most, the 1000 first patents. Those patents are then indexed more precisely.

C. Patent indexing and representation

Since patents are multilingual, or to be more precise, a language is associated to each patent's parts, we built three indexes, one per language for the patents from the targeted subset (in reality we build those indexes for the entire corpus once). In the English index, we consider the English patent parts only; in the German index, we consider the German parts only, and so on. This solution allowed us to use appropriate stop-word list and stemmer.

Each indexing term was associated with the patent part it comes from, the associated language and its frequency and becomes a variable that represents the patent (in the statistical sense). For example, the word "test", found in the abstract in English, will be counted in the variable named A_EN_test and its frequency will be kept. If it occurs in the French title too, a second variable will be defined named T_FR_test.

The indexing result can be defined as a matrix in which lines correspond to patents (individuals) observed according to the variables (indexing terms as defined previously) which are the columns of the matrix. The values inside the matrix are the term frequencies in the various parts and languages. This representation allows us to easily fuse lines or columns. Fusing lines is mandatory to fuse the various patent versions into a single one (in that case we use the mean of the frequency in each version to calculate the new term frequency). Fusing columns is useful when one wants to calculate the frequency of a term, independently to which patent parts it occurs.

IV. Analysing prior art

Prior art candidate search aims at finding patents related to the topic. It can be considered as a search problem or as a clustering problem. In this study, we chose the latter solution. In addition, the representation we chose is highly dimensional. Factorial analysis is a class of methods that aims at reducing the dimension of data whilst keeping its structure. We use Principal Component Analysis (PCA) in this study.

A. Analysing method

1) Clustering method: We chose to consider the ascending agglomerative hierarchical clustering (AHC) to group similar patents. Indeed, the target number of clusters is unknown. Unlike k-means or other methods, AHC does not require specifying the desired number of clusters. Rather, the number of clusters is chosen by looking at the graph on the decay of the node heights.

The AHC requires choosing the aggregation measure and the dissimilarity measure to use. The best aggregation measure is the Ward measure as this method makes homogeneous clusters. We use this method. With regard to the dissimilarity measure, we could use the Euclidean distance (which is the most commonly used) or other distances such as the Minkowski distance (for which the power parameter p as to be chosen) or the Manhattan distance, which are the most popular. The Manhattan distance corresponds to the norm 1, this distance will select too many patents and not the most correlated. We compared the Euclidean distance and the Minkowski distance using p=1/2 and kept the latter (the detailed results are not presented in this paper).

2) Dimension reduction: PCA is a very popular method in multidimensional statistical analysis. It makes it possible to produce graphical representations of the rows or the columns of the considered matrix, and does so in a reduced dimension space. The method is defined so that the dispersion obtained in this reduced dimension space is the largest (Jolliffe, 2002 [20]; Mardia et al., 1979 [21]). The aim of PCA is to replace a p-dimensional observation by a q linear combination of the variables, where q (the dimension of the reduced space) is much smaller than p. The linear combination defined by PCA are the eigenvectors related to the first q greatest eigenvalues.

B. Methodology

We promote a way to browse the sub-collection of patents in order to better detect prior art candidate, which is based on clustering and PCA. The method we suggest is iterative, as shown in Figure 3. First, we used a PCA in order to plot the patents and look for groups and dispersion of patents. Then, we cluster the patents in order to refine the target set of patents: the cluster containing the topic is selected. Finally, we consider the correlations between the query and the patents belonging to the cluster selected at the previous stage in order to evaluate the number of relevant patents this method selects. This methodology is applied first to patents represented by titles and abstracts, then to titles only and finally to abstracts only.

V. Results

The results we present in this section are based on the topic EP-1236420-A1 from the corpus. The experiments have been conduced on a single query topic, which is not corresponding to the CLEF-IP task. The results we are presenting here should be considered as preliminary results.

When considering the EP-1236420-A1 topic, there are 11 relevant patents associated to this topic identified as follows: EP0663161, EP0728427, EP0792603, EP0808587, EP0811336, EP0811337, EP0832580, EP0842620, EP0895734, EP1020136, EP1177745. In order to simplify the writing, we will refer to the these patents by using the

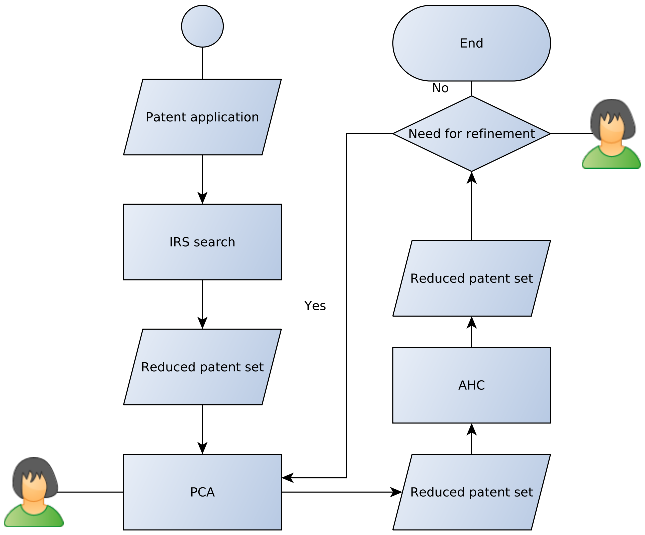

Fig. 3. Process description

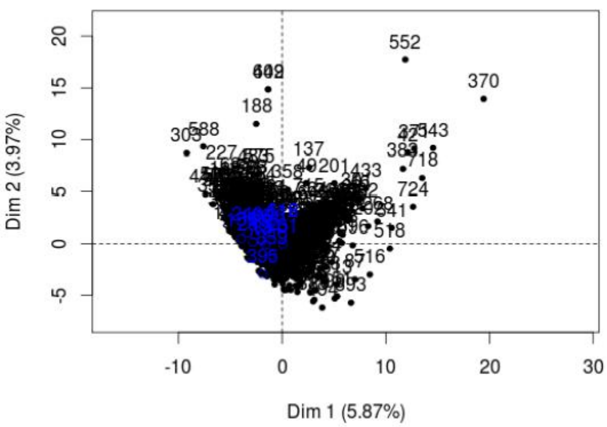

Fig. 4. Patent factor map (PCA) based on titles and abstracts using the two first dimensions

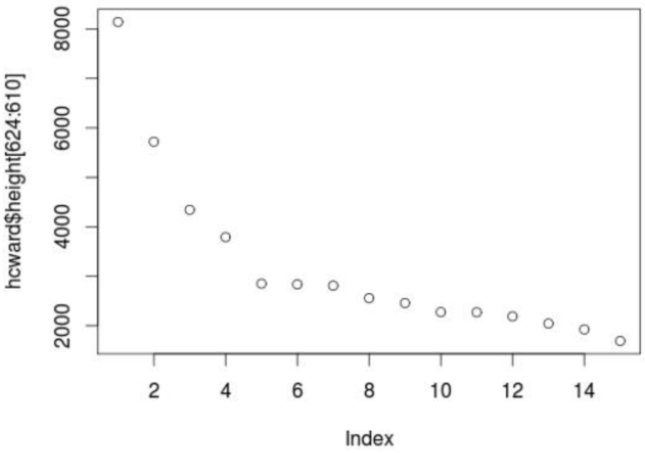

Fig. 5. Number of clusters (AHC)

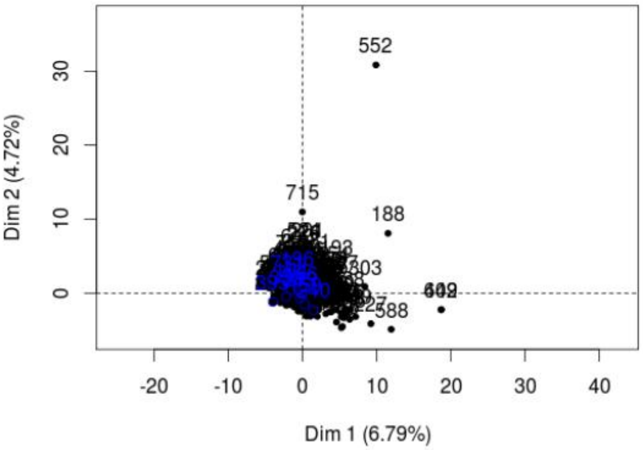

Fig. 6. Patent factor map (PCA) based on titles and abstracts on a reduce patent set

names A, B, C, D, E, F, G, H, I, J, and K respectively. The most invalidating patents are C, D, G and H.

A. English words from titles and abstracts

Figure 4 shows the representation of the patents after the first PCA. The relevant documents and the query are represented in blue, and the query is represented by the number 731. We can see three groups of patents: one in the top-left, one in the top-right and the other in the center. The query is between these three groups; from this visualization, it is difficult to know which group it belongs to. This can be explained by the fact that all these patents belong to the sub-collection related to the query 731. Moreover, we cannot identify the number of patents because the number of individuals is very large. But we can observe a ”blue” group which means that the relevant documents are close to the query. However, this graph is not very relevant to our analysis.

We will now focus on the results of the clustering which is an AHC as depicted in Section IV-A1. Figure 5 shows the graph of heights according to the number of clusters.

At this stage, we would like to obtain a sufficient number of patents in each cluster, in order to find the maximum of prior art candidates. According to Figure 2, a relevant pruning is 5 as the number of clusters. When analyzing the patent clusters, the query is in a cluster formed by 276 patents. Among these patents, we found the relevant patents labelled A, C, D, G, H, I, J and K, which means 8 out of 11.

We apply PCA on the topic cluster. Figure 6 shows the representation of the 276 patents after PCA. The query and the relevant patents are in blue; we can see one group in the center. Patents from this cluster are much related to the query. We looked to the correlations between the query and the patents belonging to the query cluster. The results are presented in the Table I. The threshold should be chosen according to the number of prior art candidates we would like to have. We can see, for example, that 0.25 is a good compromise to have a sufficient number of prior art candidates (the four most

TABLE I
RETRIEVED PATENTS ACCORDING TO QUERY-PATENT CORRELATION – TITLES AND ABSTRACTS

Threshold	Number of patents	Relevant patents
0	101	A, B,G, H, I,K
0.2	42	A,G, H, I, K
0.25	39	A,G, H, I, K
0.3	31	A,G, H, I, K
0.4	19	A,G, I, K
0.5	12	I,K
0.6	7	K
0.7	3	

TABLE II
RETRIEVED PATENTS ACCORDING TO QUERY-PATENT CORRELATION – TITLES ONLY

Threshold	Number of patents	Relevant patents
0	276	A, C, D, G, H, I, J, K
0.2	144	A, C, D, G, H, I, J
0.25	79	A, C, D, G, H, J
0.3	36	A, H, J
0.4	5	J
0.5	1	

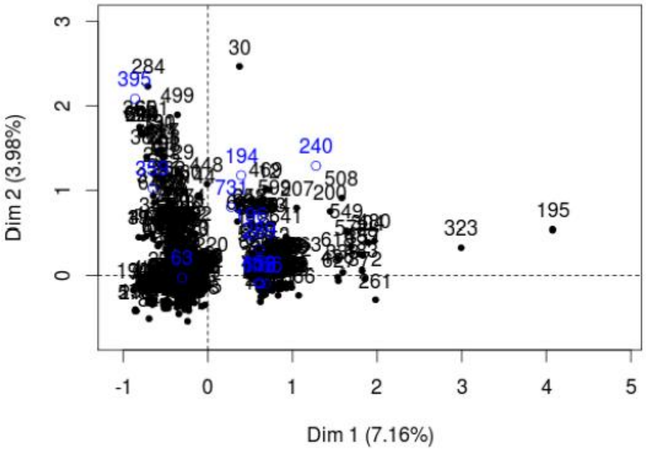

Fig. 7. Patent factor map (PCA) based on titles only using the two first dimensions

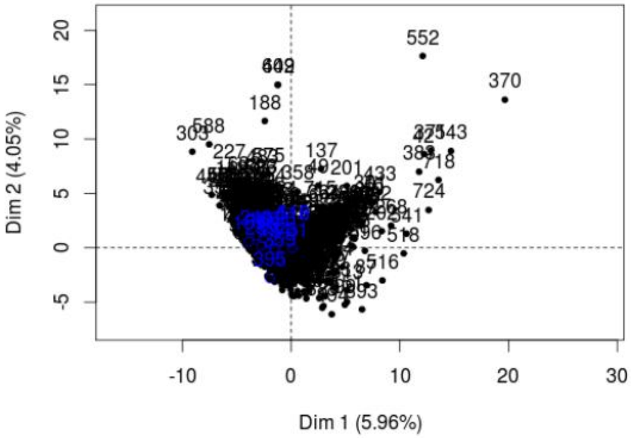

Fig. 8. Patent factor map (PCA) based on abstracts only using the two first dimensions

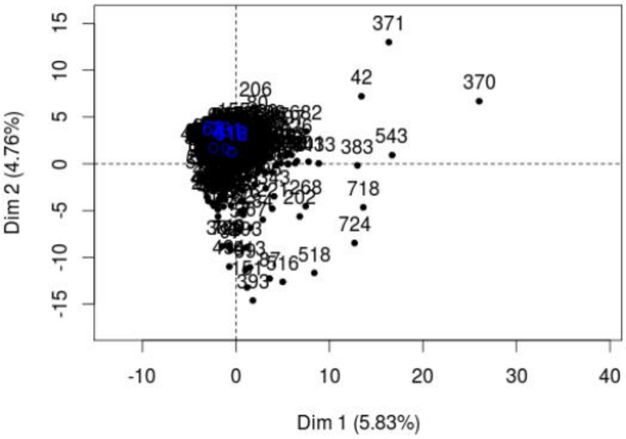

Fig. 9. Patent factor map (PCA) based on abstracts only on a reduce patent set

invalidating patents C, D, G and H are in the cluster).

B. English words from titles only

In this section, we consider the English words from the title only and reproduce the same analysis: first PCA, then a clustering on the patents and finally looking at the correlations between the query and the patents belonging to that cluster. In Figure 7, three distinct groups of patents can be observed, as well as three isolated patents. Following the method we promote, we apply clustering using AHC. The best pruning is obtained when considering 4 clusters.

The topic cluster is composed of 101 patents; it contains the relevant patents: A, B, G, H, I, K. There are 6 relevant patents out of 11. Figure 5 represents the patents after PCA was applied to the topic cluster. We can see a group of patents very close to the topic, for example the patents 359, 182, 108, 253.

Finally, we looked to the correlations between the query and the patents belonging to the query cluster. The results are presented in the Table II.

C. English words from abstracts only

The analysis is the same as previously, except that it is applied to patent abstracts only. The PCA on the sub-collection is presented in Figure 8.

We can see three groups of patents, as it was the case when considering the English words from titles and abstracts: one in the top-left, one in the top-right and the other in the center.

The query is always between these three groups so from this visualization, it is not possible to know which group it belongs to, but we still have a ”blue” group, this means that the relevant patents are highly correlated to the query.

When applying AHC, the best pruning is obtained when considering 4 clusters. The topic is in a cluster formed by 348 patents and containing the relevant patents B, E, F.

We apply a PCA on this cluster, and represent the individuals in Figure 9.

TABLE III
RETRIEVED PATENTS ACCORDING TO QUERY-PATENT CORRELATION
ABSTRACTS ONLY

Threshold	Number of patents	Relevant patents
0	348	B, E, F
0.2	208	B, E, F
0.25	115	B, E, F
0.3	58	B, E, F
0.4	10	B, E, F
0.5	1	

TABLE IV
RETRIEVED PATENTS ACCORDING TO QUERY-PATENT CORRELATION
THRESHOLD 0.25

	Number of patents	Relevant patents
Titles & Abstracts	39	A, **G**, **H**, I, K
Titles only	79	A, **C**, **D**, **G**, **H**, J
Abstracts only	115	B, E, F

The query is in the center of the group. Some patents around the group are not relevant to the query. We have lost much more relevant patents using the abstract only for clustering than when using the titles only.

When looking at the correlations between the query and the patents belonging to the query cluster we found the results presented in Table III.

VI. DISCUSSION AND CONCLUSION

The results we obtained when either considering titles and abstracts or titles or abstracts only are interesting in various ways. If we consider one particular threshold for the correlation between patents and the topic patent, patents that are obtained in the topic cluster are noticeably quite different depending on the patent part that is analyzed. Table IV presents the results for a correlation threshold of 0.25. One can see that the results obtained using title only and abstract only are disjointed. Moreover, the 4 most invalidating patents (in bold in Table IV) are retrieved and, in the case of this topic, title only is enough to find out them.

These results have been obtained using a single query patent. The process we use is iterative and interactive ; for this reason it could be difficult to compare to other methods. However, we could compare the mean average precision (or other performance measure) we obtain using our method with the values obtained by the CLEF-IP participants. That will be done in future work.

In this paper, we have shown that analysis methods can be used in an interactive way to visualize patents. This method can be used to browse a patent collection when searching for prior art candidate in an original way. As future work, we will study multilingual aspects. We think that clustering and factorial analysis could help in grouping together patents that are preliminary written in different languages thanks to the shared terms in the titles and abstracts.

REFERENCES

[1] "Derwent World Patents Index," 2012, accessed: 2013-11-10. [Online]. Available: http://www.eval-inno.eu/wiki/index.php/Derwent_World_Patents_Index

[2] "World Intellectual Property Organization," 2011, accessed: 2013-11-10. [Online]. Available: http://www.wipo.int/

[3] M. Lupu, J. Zhao, J. Huang, H. Gurulingappa, J. Fluck, M. Zimmermann, I. V. Filippov, and J. Tait, "Overview of the trec 2011 chemical ir track." in TREC. National Institute of Standards and Technology (NIST), 2011.

[4] M. Lupu, K. Mayer, J. Tait, and A. Trippe, Current Challenges in Patent Information Retrieval, ser. The Information Retrieval Series. Springer, 2011, vol. 29.

[5] J. Tait and B. Diallo, "Future patent search," in Current Challenges in Patent Information Retrieval, ser. The Information Retrieval Series. Springer, 2011, vol. 29, pp. 389–407.

[6] H. Gurulingappa, B. Müller, R. Klinger, H.-T. Mevissen, M. Hofmann-Apitius, C. M. Friedrich, and J. Fluck, "Prior art search in chemistry patents based on semantic concepts and co-citation analysis," in TREC, E. M. Voorhees and L. P. Buckland, Eds. National Institute of Standards and Technology (NIST), 2010.

[7] J. T. Florina Piroi, "CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain," IRF, Tech. Rep., 2010, tech.Rep IRF-TR-2010-00005.

[8] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz, "Clef-ip 2011: Retrieval in the intellectual property domain," in CLEF (Notebook Papers/Labs/-Workshop), V. Petras, P. Forner, and P. D. Clough, Eds., 2011.

[9] "CLEF-IP 2011 overview," 2011, accessed: 2013-09-04. [Online]. Available: http://www.ir-facility.org/clef-ip

[10] W. Magdy and G. J. F. Jones, "Applying the kiss principle for the clef-ip 2010 prior art candidate patent search task," in CLEF (Notebook Papers/LABs/Workshops), M. Braschler, D. Harman, and E. Pianta, Eds., 2010.

[11] D. Becks, T. Mandl, and C. Womser-Hacker, "Phrases or terms? the impact of different query types," in CLEF (Notebook Papers/LABs/-Workshops), M. Braschler, D. Harman, and E. Pianta, Eds., 2010.

[12] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: a cluster-based approach to browsing large document collections," in Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR '92, 1992, pp. 318–329.

[13] C. Carpineto, S. Osinski, G. Romano, and D. Weiss, "A survey of web clustering engines," ACM Computer Survey, vol. 41, no. 3, 2009.

[14] Y. G. Kim, J. H. Suh, and S. C. Park, "Visualization of patent analysis for emerging technology," Expert Systems with Applications, vol. 34, no. 3, pp. 1804–1812, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2007.01.033

[15] A. Sharma, "A survey on different text clustering techniques for patent analysis," International Journal of Engineering, 2012.

[16] S. Jun, S.-S. Park, and D.-S. Jang, "Technology forecasting using matrix map and patent clustering," Industrial Management and Data Systems, vol. 112, no. 5, pp. 786–807, 2012.

[17] S. Déjean and J. Mothe, "Visual clustering for data analysis and graphical user interfaces," in Handbook of Cluster Analysis, C. Hennig, M. Meila, F. Murtagh, and R. Rocci, Eds. http://www.crcpress.com: Chapman & Hall, CRC Press, 2014, to appear.

[18] "CLEF-IP 2011, download data, University of Technology Vienna," 2011, accessed: 2010-09-04. [Online]. Available: http://www.ifs.tuwien.ac.at/∼clef-ip/download/2011/index.shtml

[19] I. Ounis, G. Amati, P. V., B. He, C. Macdonald, and Johnson, "Terrier Information Retrieval Platform," in Proceedings of the 27th European Conference on IR Research (ECIR 2005), ser. Lecture Notes in Computer Science, vol. 3408. Springer, 2005, pp. 517–519.

[20] I. T. Jolliffe, Principal Component Analysis, 2nd ed. Springer-Verlag, 2002.

[21] K. V. Mardia, J. T. Kent, and J. M. Bibby, Multivariate Analysis. Academic Press, 1979.