UTA and SICS at CLEF-IP

Järvelin, Antti*, Järvelin, Anni* and Hansen, Preben**

* Department of Information Studies and Interactive Media
University of Tampere
{anni, antti}.jarvelin@uta.fi

** Swedish Institute of Computer Science
preben@sics.se

Abstract

University of Tampere (UTA) and Swedish Institute of Computer Science (SICS) joined forces in a CLEF-IP experiment where a simple automatic query generation approach was initially tested. For two topics, the extracted query words were compared to query keys selected by three human experts to see whether the automatic key word extraction algorithms seem to be able to extract relevant words from the topics. We participated in the main task with 8 XL runs. Despite the modesty of our results, our best runs placed relatively well compared to the other participants' runs. This suggests that our approach to the automatic query generation could be useful, and should definitely be developed further. The comparison of the queries generated by the patent engineers to those generated by our system showed that typically less than half of the query words selected by the patent engineers were present in the automatically generated queries. For the limited two topics, our automatically generated queries also performed better than the patent engineers' ones, but clearly more user data would be needed before any conclusions can be made in one way or another.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval
Keywords: Patent retrieval, automatic query generation, user assessments

1. Introduction

Patents are a valuable source of scientific and technological information. However, patent retrieval is a challenging task and identifying relevant patent documents among millions of international patents is time consuming, even for professional patent officers and especially for laymen applying for patents. Therefore there is a clear need for effective patent information retrieval systems. The importance of effective automatic methods for patent retrieval is increasing in global society, where also the patents granted in other countries may need to be found and systems capable of crossing language barriers are needed.

Therefore we were interested in the CLEF-IP track. As this was our first try with patent retrieval, we just concentrated on getting a retrieval system up and running and on studying the automatic query generation process. We participated in the main task of the CLEF-IP track (for overview, see Roda et al. 2009) with 8 XL runs, where different query word selection tactics were tested. We felt that the best way to analyze the performance of the automatic query word selection would be a comparison against the query word selection of human experts. Patent retrieval experts are hard to come by, but we had the opportunity to have three experts to analyze two of the topic documents. This allowed us to get a glimpse of what words and from which fields of the patent should be selected for the queries. The experts were also asked to write down other keywords or phrases they would use when searching for prior art for each of the topic documents. This way we could analyze (for the limited two topics) the need of outside sources of expansion terms to the queries. We ran the queries produced in the manual analysis of the topics in the CLEF-IP collection to get some performance data to compare with the automatically generated queries' performance.

The paper is organized as follows: our approaches to the automatic query generation are presented in Section 2 together with a short description of the manual topic analysis and a description of the runs. Section 3 then describes the technical details of our retrieval system. The results are presented in Section 4 and Section 5 concludes with a short discussion.

2. Query generation

2.1 Automatic query generation

One of the questions that need to be solved for automatic patent retrieval to be feasible is how to automatically generate queries of patents or patent applications. This is a complex task and usually carried out by patent engineers possessing extensive knowledge of both the domain of the invention and of the patent text genre. The complexity arises mostly from the language used in patent documents – a mixture of legislative language and domain specific technical sublanguages. The technical fields are various and the writing styles and sublanguages used in patents differ notably, which effects the term-extraction and weighting algorithms in information retrieval systems (Mase et al. 2005). While natural language texts always present a vocabulary mismatch problem for information retrieval, the problem is compounded in patent retrieval by the frequent use of novel vocabulary: virtually every patent uses new words and the situation is made worse by the intentional use of nonstandard or vague terminology by some inventors (Larkey 1999). The idiosyncratic legal styles and terminology can lead to spurious similarities between patents based on style rather than content (Larkey 1999).

Our intention in CLEF-IP was to make some preliminary tests concerning which fields of the patent documents, and which combinations of the fields would be the best sources of query keys, as well as to test two simple approaches to query key extraction from the topics. European patent applications include several fields with rather strictly defined contents, of which especially the text fields title, abstract, description and claims are important sources of query keys. None of these fields stands out as the clearly preferable source of query keys: The claims include the most central content of a patent, i.e., the exact definition of what the invention covered by a patent is and have therefore often been used as the primary source of query keys in patent retrieval studies (Mase et al. 2005; Fujii et al. 2004). The claims however, are written in a legislative language (unlike the other fields), which may lead to spurious similarities between patents based on style rather than content (Larkey 1999). Titles and abstracts on the other hand, are often written to be uninformative, while descriptions are written to be as wide as possible and may thus contain words too general to be used as query keys. Therefore, we tested several combinations of the title, abstract, description and claims fields as sources of query keys. Only the combinations that turned out to be the best ones during training were used in the final CLEF-IP run and are reported here.

Query keys were extracted separately from the topics for all of the topic languages and three monolingual queries were formed. The main languages of the topic documents were recognized. The main language topic fields were not processed further, but, in case a field that was to be used as a source of query keys lacked content in any of the other two languages, Google Translate was used to translate the missing field from the main language of the topic. We used Java-API available at http://code.google.com/p/google-api-translate-java/ for the Google Translate. The API restricts the length of the translated text to approximately 1,500 bytes. Therefore we often could not translate the whole field, but needed to preselect the source language sentences so that the 1,500 bytes limit would not exceed. Thus we selected the sentences to be translated with the following procedure:

1. The sentences were ranked according to their average RATF-scores (Relative Average Term Frequneces) of their words. The RATF score (Pirkola et al. 2002) of a word is calculated with

$$RATF(k) = 10^3 \frac{cf_k / df_k}{\left(\ln\left(df_k + SP\right)\right)^P},$$

where cf_k is the collection frequency of word k, df_k its document frequency, and SP and P collection depended parameters. In this study we used SP=3,000, and P=3. The training set suggested, that the exact values of parameters SP and P did not matter that much.

2. The best sentences were then selected until the upper limit of 1,500 bytes was reached.
3. The selected sentences were translated using the Google Translate Java-API, and the translated text was added to the missing field of the target language's patent.

We had two different approaches to automatic extraction of query words: using the standard *tf·idf* weighting and using the RATF scores modified slightly to better suit the retrieval scenario. While RATF is originally based on using collection frequencies only, we modified the formula to include also the word frequencies in the topics. The modified RATF score for a key k is computed as follows:

$$\begin{aligned} RATF_{mod}(k) &= \frac{tf_k}{cf_k/df_k} RATF(k) \\ &= \frac{tf_k}{cf_k/df_k}\left(10^3 \frac{cf_k/df_k}{\left(\ln(df_k + SP)\right)^P}\right), \\ &= 10^3 \frac{tf_k}{\left(\ln(df_k + SP)\right)^P} \end{aligned}$$

where tf_k is the frequency of the key k in a topic document, cf_k its frequency in the collection, and df_k the number of documents in which the key k occurs. Thus the tf_k is normalized by dividing it by the logarithm of the document frequency of key k scaled with the collection depended parameters SP and P. Again, the training set suggested that the values of parameters SP and P had only minimal effect to the results.

Also, a separate query was formed of the IPC codes present at the topic document. We did extensive training runs before the official runs and found that the IPC codes were not always useful in our system, but sometimes even damaged the performance of some runs. Therefore in some cases, we did not include the IPC run in the final runs. We ran the four queries separately, the monolingual queries to the corresponding monolingual indices and the IPC query to a separate index containing only the codes. The results were merged in a later phase. The following variants were tested in our eight CLEF-IP runs:

- All fields (title, abstract, claims, and description), with IPC, RATF and translation (UTASICS_all-ratf-GT-ipcr_Main_XL)

- All fields (title, abstract, claims, and description), with IPC, RATF and no translation (UTASICS_all-ratf-ipcr_Main_XL)

- All fields (title, abstract, claims, and description), with IPC, *tf·idf* and translation (UTASICS_all-tf-idf-GT-ipcr_Main_XL)

- All fields (title, abstract, claims, and description), with IPC, *tf·idf* and no translation (UTASICS_all-tf-idf-ipcr_Main_XL)

- Abstract and description, with RATF and translation. No IPC! (UTASICS_abs-des-ratf-GT_Main_XL)

- Abstract and description, with RATF and no translation. No IPC! (UTASICS_abs-des-ratf_Main_XL)

- Abstract title, and claims, with IPC, RATF and translation (UTASICS_abs-tit-cla-ratf-GT-ipcr_Main_XL)

- Abstract, title and claims, with IPC, RATF and no translation (UTASICS_abs-tit-cla-ratf-ipcr_Main_XL)

2.2 manual query generation

In addition to the automatically generated queries, we also wanted a set of user generated keywords and queries on the same topics in order to compare to the performance of the automatically generated queries. As patent examiners are hard to come by, it was clear that the set of topics that could be analyzed was very limited. We ended up with having three patent examiners to examine three topics. All examiners were asked to analyze the same three patents, to get an idea of how much their analyses differ from each other. The goal with this subtask of the experiment was to repeat a real-life initial patent query formulation situation as accurately as possible, as an answer to a problem identification and information need. Three professional patent engineers from a Swedish patent bureau collaborated with us on this task. The purpose was to examine how:

a) Initial assessments was done,

b) What type of keywords and other features of the patent text (such as classification codes, document numbers, synonyms, etc) was used and

c) A first initial query could be formulated.

For this subtask, three patent topics from the CLEF-IP patent topic set were selected: EP1186311, EP1353525, EP1372507. Due to time constraints, only the two first of these could be analyzed by now and thus the results are only reported for the two first topics. We are aware that not much can be claimed based on an analysis two topics only (or three for that matter), but were hoping to get some first indications on how patent examiners work and our automatic query generation procedure could be improved.

The manual analysis of the topics contained three different stages. a) Pre-questionnaire collection data of preferences of patent engineers, b) individual manual topics analysis; and c) post-questionnaire, including data on the task performed and keywords generated. First the three patent engineers were introduced to the patent task as well as the purpose of the study. They were also introduced to the CLEF-IP topics and what was expected of them. All three agreed upon the study set-up. Before the data collection started, each patent engineer received a paper copy of the three topics, paper copies of the pre-questionnaire and a post-questionnaire. Along these documents, there were instructions on how to mark up the selected words and additional comments related to the query generation. The patent engineers were asked to perform the assessments of the topics as close to a real situation as possible.

The individual manual analysis of the patent applications were done as follows: First the three professional patent engineers were asked to read a full version of a copy of the CLIF-IP topic, i.e., a patent application. Then, the patent engineers were asked to write down, mark or by other means point out relevant items in the patent application that described and represented the problem and need for information. These items could be keywords, terms, classification codes, document number, synonyms or relevant sections and could be chosen from any of the patent fields. After selecting relevant words from the patent applications, the patent examiners were also asked to write down synonyms to the selected words or other keywords that they might find useful when searching for prior art. Finally, the patent engineers were asked to sort the different items into to five different classes:

1. Top5 keywords
2. Top5 keywords + synonyms
3. Top10 keywords
4. Top10 keywords + synonyms (with top5 ranked)
5. Initial query that the patent examiner would use for starting to search for prior art concerning the topic. This could contain any number of words, phrases and classification codes and the examiners were allowed to define query structure.

For the CLEF-IP experiments, we formed manual runs from the top10 keywords and from the top10 keywords + synonyms. Separate queries were formed of the IPC codes, corresponding to the approach the automatic runs. These were compared to corresponding runs using automatic query generation with 10 top ranked keywords.

3. System details

We used the Indri engine of the Lemur toolkit (Strohman et al. 2005) for indexing and retrieval. Our approach to indexing was to generate a single "virtual patent" of the several versions that may exist for each patent, in a manner similar to the topic generation by the organizers: only the central fields (title, abstract, description and claims) were indexed and only the most recent version of each of the fields was indexed. We created a separate monolingual index for each of the languages (one English, one French and one German index). The content in the different languages was separated based on the language tags (present in the patent XML code) and a "virtual patent" was created for each language. The IPC codes were indexed separately into a language independent index. This way we created in total four indices that could be searched separately. The natural language index words were stemmed using the popular Snowball stemmer for all of the languages. The IPC codes were truncated after the fourth character.

As a consequence of indexing the content in each of the languages separately, we needed to run three monolingual queries and an IPC code query for each of the topics. All the natural language queries in all runs were set to include 50 words, based on training results. The IPC queries included all the IPC codes present in a topic document. The results from the four different queries were merged at query time separately for each topic using MAD (Mean Average Distance) merging model developed by Wilkins and colleagues (2006). MAD is a model for relative weight determination based on the idea that the score distribution in the result lists for well performing queries (indices) will differ from the score distributions typical for bad queries (indices). Good performance typically correlates with a high degree of change in document scores at the beginning of a result list, while poor performance results in linearly decreasing scores (Wilkins et al. 2006). MAD is computed by summing for all documents in a result list the difference in score between the document at rank *n* and the document at rank *n*+1. The sum is then divided by *N*-1, *N* being the total number of documents in the result list:

$$MAD = \frac{\sum_{n=1}^{N} \left(score(n) - score(n+1) \right)}{N-1} .$$

To achieve scores that are comparable between result lists, a ratio of MAD within a top subset of a result list versus that of a larger set of the same result list is needed. This score is called similarity cluster (*SC*) and is calculated as

$$SC = \frac{MAD(subset)}{MAD(larger\ subset)} .$$

Then the final weight for each feature (index) is the relative percentage of the total a given *SC* score is against the sum of all *SC* scores

$$f\,eatureweight = \frac{Feature\ SC}{\sum all\ SC\ scores} .$$

Before calculating the MAD scores the results from the indices were normalized to range [0-1] using the min-max normalization. Each index returned at most top 2,000 documents matching to the query. The size of the smaller subset was 10 % of the returned documents, and the size of the larger subset was 90 % of the returned documents. After merging top 1,000 documents were returned as a result of the query.

4. Results

We prefer the use of nDCG metric for evaluation, as it is a user oriented metric. Nevertheless, we will also report MAP and precision at 10 retrieved documents (P10) in Table 1, as these metrics are commonly used. The nDCG results for highly relevant documents are excluded, as they were exactly the same as for all relevant documents. Anyhow, all of the three metrics gave approximately the same evaluation results: our best run was always the one generated from the abstract and description fields only using the (modified) RATF formula and without translation (UTASICS_abs-des-ratf). The worst one was the otherwise identical run, but where Google Translate (GT) was used for query translation (UTASICS_abs-des-ratf-GT). The rest of our runs had very similar performances and there

Run ID	P10	MAP	nDCG
UTASICS_abs-des-ratf	**0.0945**	**0.1237**	**0.4722**
UTASICS_abs-des-ratf-GT	0.0779	0.0996	0.4375
UTASICS_abs-tit-cla-ratf-ipcr	0.0838	0.1088	0.4405
UTASICS_abs-tit-cla-ratf-GT-ipcr	0.0848	0.1097	0.4416
UTASICS_all-ratf-ipcr	0.0923	0.1209	0.4507
UTASICS_all-ratf-GT-ipcr	0.0934	0.1227	0.4511
UTASICS_all-tf-idf-ipcr	0.0930	0.1216	0.4562
UTASICS_all-tf-idf-GT-ipcr	0.0934	0.1217	0.4518
humb_1	*0.1776*	*0.2802*	*0.5877*

Table 1. Results for the XL runs. Our eight runs compared to the best run by the Humboldt University, humb_1.

were only small changes in the order of the other runs between the evaluation measures. Despite the modesty of our results, our best runs placed relatively well compared to the other participants' runs. It should nevertheless be noted that the run (humb_1) by the Humboldt University (also given in Table 1) totally outclassed all others and that our best run, in terms of MAP, did not reach even 50 per cent of their performance.

Our best (and worst) run differed from the rest of our runs by three factors: fewer fields were used as sources for query keys, the IPC codes were not used and the translation using GT was clearly harmful. The reason for not using IPC codes in these runs was that our training runs showed that adding the IPC query to this type of runs only caused the performance to deteriorate. This was not the case with the other types of runs. Therefore it is difficult to claim that the IPC codes (as used in our runs) were not useful in general. The combination of the abstract and description fields seemed to be a better source of query keys than the other combinations. Abstracts (used in all of our runs) in general were the most promising source of query keys when no proper translation resources were available: all topics contained the abstracts in all of the three target languages. Using GT was not useful in general and seemed to perform especially badly on translation of the description fields. RATF and *tf·idf* performed very similarly, and based on our results both of them could be used for automatic query generation.

The Tables 2 and 3 present the overlaps between the automatically generated ten-word queries and queries generated from the words that the patent engineers selected to be the top 10 representative words for the topics EP1186311 and EP1353525. Sometimes the engineers had selected phrases or compounds, such as "optical network", instead of single words. In these cases all constituting words of a phrase or compound were included into the query. This also introduced duplicates into the queries, e.g., with pair "optical cross-connect" and "optical network", where word "optical" occurs twice. Such duplicates were removed, but still many user-generated queries contained more than ten words. All queries were stemmed and stopped before their evaluation. For the topic EP1186311 (Table 2) the user A had only selected eight words into the top 10 list.

The overlap between the user-generated and the automatically generated queries is usually four words. Therefore the users and our system tended to select relatively different words for top 10 queries. Interestingly, the user generated queries performed worse than the automatically generated ones. The MAP for automatically generated queries was 0.3333 for the topic EP1186311 and 0.0004 for the topic EP1353525. All user-generated queries for the topic EP1186311 had MAP clearly less than 0.01. The MAP for the topic EP1353525 was 0 for all user-generated queries. We also generated "ideal queries" containing all unique query words selected by users A-C, but these queries did not perform any better. Expanding the user-generated queries with the synonyms selected by the users did not change the situation either. However, it has to be noted that more user data would probably change the situation quite a bit; at least the performance gap between the automatic and the user-generated queries would definitely shrink.

USER A	
Auto Q	plunger syring pump leadscrew obstruct motor actuat head encod speed
User Q	a61m syring motor speed rotat stop alarm obstruct
Overlap	4
USER B	
Auto Q	plunger syring pump leadscrew obstruct motor actuat head encod speed
User Q	a61m syring pump alarm obstruct rotat speed optic sensor detect
Overlap	4
USER C	
Auto Q	plunger syring pump leadscrew obstruct motor actuat head encod speed
User Q	a61m syring pump motor obstruct speed rotat slow detect fall
Overlap	5

Table 2. The overlap between the queries generated from the top 10 words selected by the patent engineers and the automatically generated ten-word queries for the topic EP1186311. “Auto Q” refers to the automatically generated queries and “User Q” to the user-generated queries.

USER A	
Auto Q	optic turnaround oxc node signal transmiss mcu transmitt receiv transmit
User Q	h04q optic network transmiss cross-connect oxc rate stabil node
Overlap	4
USER B	
Auto Q	optic turnaround oxc node signal transmiss mcu transmitt receiv transmit
User Q	optic cross-connect oxc shorten time h04q 11 network node transmiss rate speed plural line
Overlap	4
USER C	
Auto Q	optic turnaround oxc node signal transmiss mcu transmitt receiv transmit
User Q	optic cross-connect h04q transmiss speed rate adjac determin node neighbour switch
Overlap	3

Table 3. The overlap between the queries generated from the top 10 words selected by the patent engineers and the automatically generated ten-word queries for the topic EP1353525. “Auto Q” refers to the automatically generated queries and “User Q” to the user-generated queries.

5. Discussion

Based on the results submitted by the CLEF-IP participants we can be happy to the performance level of our system. Although our results were modest, our system performed quite well compared to the other participants' systems. Thus our results give us a good starting point in exploring the field of patent retrieval further. Our results also suggest that the modified version of the RATF-formula and the more traditional *tf·idf* weighting could be viable alternatives for automatic query word extraction in the patent retrieval. Based on our results it seems that the combination of abstract and description fields of the patents are the best sources for picking up relevant query words from the patents – at least this was the case with our approach to the automatic query generation.

However, automatic query generation is difficult due to the fact that the good query words are often not present in the patent documents. Therefore an outside source of synonyms and expansion terms for the queries should be used. Our approach of using the Google Translate for translating the missing patent fields did not noticeably improve the results. For some runs the results even deteriorated. A proper translation resource should greatly improve the results.

We indexed the patents of each language as well as the IPC-codes into separate indices. This required us to use post-query result merging to produce a single ranked result list. The good side of this solution is that it is easy to run queries also in only one or two of the three languages. But this approach complicates the system and the effects on the final ranking are not known. It might also be a better choice to index the IPC-codes directly to the language indices. However, we did not have resources (time) to test this approach. Furthermore we truncated the IPC-codes after the fourth character, and the effects of this procedure should be further investigated.

The results from the manual analysis showed mostly that we would need more topics to be able to say something reasonable from these results. The two topics used were randomly selected without knowledge of their performance – a choice that resulted with two topics with rather untypical performance (based on the MAPs of the queries). Therefore these results should be treated at most as tentative. However, it is clear that if more user data could be obtained, interesting evaluations of the system could be performed. This would also give valuable information on how our system could be enhanced, especially from the automatic query generation's point of view.

Acknowledgements

The work of Anni Järvelin is funded by Tampere Graduate School in Information Science and Engineering (TISE), and Antti Järvelin was partly funded by the Finnish Cultural Foundation. We wish to thank the Swedish patent bureau Uppdragshuset for their kind help and especially the three patent examiners for participating in the user assessments.

References

Fujii, A., Iwayama, M., and Kando, N. (2004): Overview of Patent Retrieval Task at NTCIR-4. In: *Proceedings of the NTCIR-4*, Tokyo, April 2003-June 2004.

Larkey, L. (1999): A Patent Search and Classification System. In: *Digital Libraries 99 - The Fourth ACM Conference on Digital Libraries* (Berkeley, CA, Aug. 11-14 1999). ACM Press: pp. 79-87.

Mase, H., Matsubayashi, T., Ogawa, Y., and Iwayama, M. (2005): Proposal of Two-Stage Patent Retrieval Method Considering the Claim Structure. ACM Transactions on Asian Language Information Processing, 4(2): pp. 186-202.

Pirkola, A, Leppänen, E., and Järvelin, K. (2002): The RATF formula (Kwok's formula): Exploiting average term frequency in cross-language retrieval. Information Research, 7(2).

Roda, G., Tait, J., Piroi, F., Zenz, V. (2009): CLEF-IP 2009: Retrieval experiments in the Intellectual Property domain. In: *CLEF working notes 2009*, Corfu, Greece, 2009.

Strohman, T., Metzler, D., Turtle, H., Croft, W.B. (2005): Indri: A language-model based search engine for complex queries. In: *Proceedings of the International Conference on Intelligence Analysis*.

Wilkins, P., Ferguson, P., Smeaton., A.F. (2006): Using score distributions for query-time fusion in multimedia retrieval. In: *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*. ACM Press: pp. 51-60.