

BiTeM site report for TREC Chemistry 2010: Impact of Citations Feedback for Patent Prior Art Search and Chemical Compounds Expansion for Ad Hoc Retrieval

J. Gobeill^a, A. Gaudinat^a, E. Pasche^b, D. Teodoro^b, D. Vishnyakova^b, P. Ruch^a

^a*BiTeM group, University of Applied Sciences, Information Studies, Geneva*

^b*BiTeM group, University and Hospitals of Geneva, Geneva*

contact: {julien.gobeill;patrick.ruch}@hesge.ch

Abstract

For two years, the TREC Chemical Track aims at evaluating participant systems in chemical patent searching. In 2010, it continued with the two tasks from 2009: Prior Art search (PA) and Technology Survey (TS). The BiTeM group participated in both tasks and obtained satisfactory results, relying on a large panel of strategies which were evaluated within the framework of past similar competitions. There are three main conclusions that we draw from this campaign. First of all, regarding a baseline computed by Information Retrieval (IR) only, simplest models achieved the best results for both tasks, such as indexing only titles, abstracts, and claims, and no stemming; however, for the PA task, the performance of this baseline remains low (Mean Average Precision 0.043) compared to last year (MAP 0.067). Further analysis of the query set reveals that description sections were in 2010 twice longer than in 2009, while citations number was stable; having longer queries obviously resulted in a degradation of the signal-to-noise ratio, and in a more complex task for standard IR. Secondly, IPC codes were of no use for the PA task, and even decreased performances, whether they were injected in the index or used for filtering the results. Because this strategy is effective when applied to EPO patents in general domain, further experiments or expertise need to determine if it fails because it is applied to a specific domain, or because the quality of IPC annotations in USPTO patents is insufficient. The last conclusion deals with our re-ranking strategy based on citations feedback for the PA task. Such a strategy led to a dramatic improvement from 0.043 to 0.261 for MAP (+ 507%), and from 0.31 to 0.62 for Recall at 500 (+ 100%). Further analysis shows that our citations feedback strategy achieves to strongly capture the chemical applicants' behaviour, which tends to cite regular patterns of multiple patents massively inter-connected with direct citations. Results of the TS task prove the effectiveness of synonyms expansion driven by chemical entities normalization.

Introduction

Since 2009, the TREC Chemical (TREC-CHEM) Track focuses on evaluation of search techniques for discovery of digitally stored information in chemical patents and academic journal articles [1]. In 2010, it continued with the two tasks from 2009: Prior Art search (PA) and Technology Survey (TS) [2]. In the PA task, participants' systems had to find relevant patents with respect to a query set of 1'000 patents (333 from USPTO, 334 from EPO, and 333 from WIPO). The evaluation was based on existing citations of these patents. In the TS task, participants systems had to retrieve relevant documents with respect to a set of 28

topics provided by chemical patent experts; the evaluation was based on human judgements. For both tasks, the collection was very similar to the past TREC-CHEM Track, and contained approximately 1'320'000 patents from EPO (10%), USPTO (70%) and WIPO (20%). For the TS task, the collection besides contained approximately 181'000 scientific articles from many publishers.

BiTeM (Bibliomics and Text Mining) is a research group located in Geneva, having a strong expertise on text mining in large corpora, especially in biomedicine [3]. In 2010, we participated in both tasks of TREC-CHEM, relying on the background we have acquired during past similar competitions, such as TREC-CHEM 2009, CLEF-IP 2009 and 2010 [4,5] and PatOlympics

[6]. The CLEF-IP competitions contained a similar Prior Art search task, with the difference that patents were all issued from EPO and were not domain-specific. The PatOlympics competition was similar to the TS task, with the Chemathlon task, also focused on chemistry, but the systems were evaluated by live interaction with patent experts. In this paper, we will sometimes not detail some technical issues and will simply refer to our past publications within these competitions [7,8,9].

Data

While TREC-CHEM 2009 and 2010 data, i.e. the collection and the query sets, look very similar, there are actually notable differences, especially for the PA query set.

1) Collection

Foremost, the TREC-CHEM 2010 collection contained patent documents, sometimes referring to a same patent (such as “EP-0218350-A1” and “EP-0218350-B1”). As for past competitions, we decided to merge all documents belonging to the same patent into a unique and virtual patent file. As evaluations were conducted at the document level, when a virtual patent file was considered as relevant, we simply considered all its documents as relevant. Parsing scripts were run in order to extract several sections from the documents: *Title*, *Abstract*, *Claims*, *Description*, *Applicants*, *Inventors*, and *IPC codes* were thus extracted in order to be injected in the Information Retrieval (IR) model, while

IPC codes, *Application Dates*, *Publication Dates* and *Citations* were extracted in order to be exploited as metadata for post-processing strategies. *IPC codes* were extracted at different levels: subclasses (e.g. A61K) and subgroups (e.g. A61K 8/00).

The TREC-CHEM collection also contained scientific papers for the TS task. Parsing scripts were run in order to extract *Title*, *Abstract* and *Body* sections. Papers were then added to the index for the TS task.

2) Query sets

The 2010 query set for the PA task contained 1’000 patents, as the 2009 one. Moreover, posterior analysis showed that the average number of citations to find was 43, against 44 in 2009. But the average size of patents *Description* sections was actually more than twice longer in 2010: 13’000 words in 2010 versus 5’700 in 2009. Figure 1 shows that this difference was present all along the set. The difference was less significant between the average lengths of *Claims* section: 1’330 words in 2010 versus 1’210 in 2009. Moreover, for the 2010 query set, the description length was quite correlated with the number of citations (Pearson correlation coefficient 0.21), while the claims length was not (0.07). In the rest of this paper, we will make several assumptions from these facts.

Because such differences were ignored when we designed our system, we tuned it with the 2009 query sets for the PA and TS tasks. Parsing scripts were run on the training and the query sets in order to extract the same sections as for the collection.

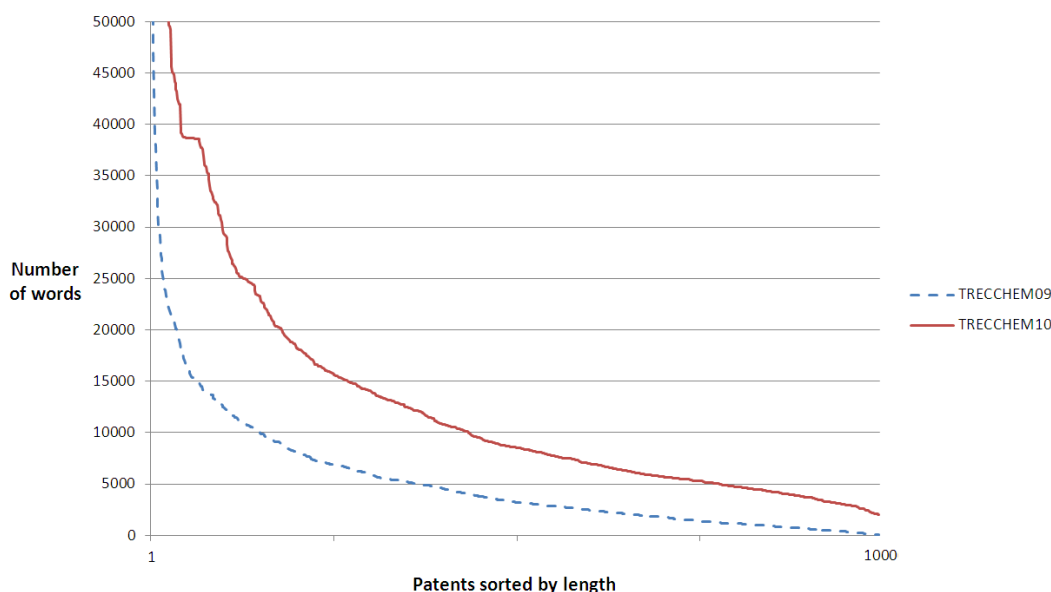


Figure 1. Comparison of the number of words in patents description sections along the PA task query set, between 2009 and 2010.

Strategies

Several strategies were investigated during the three steps of our system's pipeline (the pre-processing step, the Information Retrieval itself, and the post-processing step). As for the official evaluation, Mean Average Precision (MAP) was computed in order to evaluate the different strategies and settings. We also chose to consider Recall at 500. As the average citations number to find is about 40, we think that 500 is a realistic threshold in order to interpret the recall.

1) Document representation

Patent documents are often large files, containing much information structured into several sections. Thanks to this structure, we easily could extract the information from each section, and investigated different document representations, in order to determine which one achieved the best retrieval results. Therefore, we decided to start from a document representation similar to the one that achieved the best results in TREC-CHEM 2009, and which simply consisted of *Titles*, *Abstract* and *Claims*; we then computed a first run with a default IR model. From this, we were able to inject the above-mentioned other sections and to evaluate their contribution in the system.

2) Information Retrieval

Indexing and Retrieval were computed with the Terrier platform [10], which is designed for large collections and which we already used in past similar competitions. We chose settings which proved to be efficient in the past PatOlympics competition: PL2 as weighting scheme, and Bo1 as Query Expansion model, both with default parameters and without stemming. Once found the best document representation, we also investigated other promising weighting schemes implemented in Terrier, such as Okapi-BM25; see [11] for more details about algorithms implemented in Terrier. The contribution of stemming (Porter [12]) was also investigated.

3) Re-ranking based on citations feedback

One post-processing strategy that achieved spectacular improvement of our performances in TREC-CHEM 2009 PA task was re-ranking based on citations feedback (+ 168 % for MAP in 2009). Using such a simple approach, the run was ranked #1 in 2009. This strategy consists in promoting patents that are cited by the retrieved patents, regarding the retrieval status values and weighted by a constant α . It enables to

retrieve some relevant patents which are impossible to retrieve with a standard IR approach, because they sometimes share only few words with the query. See [6] for more technical details.

4) Filtering based on IPC codes

Another post-processing strategy that we largely investigated in past similar PA tasks is filtering based on IPC codes. This strategy consists in removing retrieved patents that do not share any IPC code with the query. This strategy proved to be efficient in the past CLEF-IP tracks, while it was of no use in TREC-CHEM 2009. In 2010, this strategy was again investigated, at the levels of subclasses and subgroups.

5) Filtering based on Publication Dates

For the PA task, as a patent obviously cannot be cited before its Publication Date, we applied a simple strategy in order to remove all patents in which the Publication Date was posterior to the Application Date of the query.

6) Query Expansion using chemical annotation for the TS task

Despite the limited impact of chemical expansion on our performances in TREC-CHEM 2009 TS task (+5%), we decided to continue further investigations with such a strategy [7]. The first step of the chemical expansion consisted to identify the chemical terms in patents; this is achieved using Oscar3 tool [13], which detect the boundaries of chemical terms and assign a confidence score to each detected term. This score is based on the term itself, but also on the context in which the term has been found. The next steps consists to normalize the identified entities using the MeSH categorizer and finally to extract synonyms from different chemical databases, such as PubChem. We dedicated three runs to this strategy. Two runs uses only the chemical terms identified with a confidence score higher to 0.8; the first run (small QE) used only the main term while the second run (medium QE) used both the main term and the set of synonyms. The third run (large QE) used all chemical terms identified and their set of synonyms. Our engine for tagging chemistry entries in text is available online – ChemTagger [21].

7) Query Expansion using IPC automatic categorizer for the TS task

Similarly to the PA, IPC codes are also very important to improve searching performance in the TS tasks [7]. However, in this case they were not available at the searching time. We used an automatic IPC classifier –

IPCCat [20] – to assign codes to TS requests [9,16]. Then, we injected them into the model using both formats subclass and subgroup.

Results and Discussion

For the PA task, we investigated and evaluated the different strategies and settings with the training set, and finally submitted one official run. We submitted several runs for the TS task in order to test different Query Expansion strategies.

1) Document Representation

Evaluated with the 2010 gold file provided by the organizers for the PA task, the performance of the Document Representation baseline is 0.043 for MAP and 0.31 for R500. Moreover, further experiments with the 2010 gold file showed that the other above-mentioned sections were of no use when they were injected in the model (Table 1).

Document representation	MAP
Baseline	0.043
Baseline + <i>Inventors & Applicants</i>	0.038 (-11%)
Baseline + <i>Description</i>	0.042 (-1%)
Baseline + <i>IPC codes (subclasses)</i>	0.04 (-6%)
Baseline + <i>IPC codes (subgroups)</i>	0.038 (-12%)

Table 1. Results for different Document Representations.

In TREC-CHEM 2009, the MAP at this step was 0.067. While there was more material in the queries, the 2010 longer patents obviously resulted in a degradation of

the signal-to-noise ratio, and in a more complex task for IR. Another observation is the strong degradation led by IPC codes (-6% for subclasses, -12% for subgroups), while they significantly improved the performances (+3% for subclasses, +8% for subgroups) in CLEF-IP tracks [8]. Further conclusions about IPC codes will be drawn in the Filtering based on IPC codes section.

2) Information Retrieval

The baseline model used for investigating the Document Representation (PL2 weighting scheme, Bo1 Query Expansion) was finally the best one for the system. Moreover, Porter stemming degrades the performances (-8% for MAP); this result is coherent with similar experiments in Chemathlon, and tends to prove that stemming is of no use for such chemical collections.

3) Re-ranking based on citations feedback

In TREC-CHEM 2010, this strategy showed a further improvement than in 2009: MAP reached from 0.043 to 0.261 (+507%) and R500 reached from 0.31 to 0.62 (+100%). The best value for α was this time 0.3. In the CLEF-IP tracks, this strategy only achieved a slight improvement (+3%) [7,8].

In order to interpret this result, we decided to split the query set in four equal parts, regarding their description length. Thus, part 1 consisted of the 250 shortest patents (average word number 3'860) while part 4 consisted of the 250 longest (average word number 33'000). We then compared the performances obtained on these four sets, before the re-ranking based on citations feedback (i.e. for IR only: IR column) and after it (IR + CitFB column). Results are presented in Figure 2.






	average length	average citations		IR (MAP)	IR + CitFB (MAP)
part 1	3860	32.5		0.045	0.208
part 2	6740	36.2		0.048	0.238
part 3	11800	47		0.048	0.282
part 4	33000	53.7		0.032	0.335
all the query set	13800	42.3		0.043	0.261

Figure 2. Comparison of the performances before (IR column) and after (IR + CitFB column) the citations feedback, regarding patents lengths.

First of all, we observe that the longer the patent is, the more patents it cites (average citations column). Such a result is consistent with the correlation coefficient presented above. Then, we observe that the IR alone is less efficient in the fourth part (IR column); such a result consolidates the assumption that in these

chemical patents, long descriptions degrade the signal-to-noise ratio, and are a challenging issue for standard IR. On the contrary, we observe that the longer the patents are, the more powerful the citations feedback strategy (IR+CitFB column) is, up to a +1000% improvement for the fourth part. Because such a

strategy is less effective when applied to EPO patents in general field, we can make the assumption that it is an effect of the restricted domain (chemistry) or an effect of the origin (patent offices) of patents.

Patent office	USPTO	WIPO	EPO
Number of patents in the topic set	333	333	334
Average number of citations to find	66	35	25
Inter-connection of cited patents	70%	59%	57%

Table 2. Statistics on topic set’s patents regarding their office.

In the overview of TREC-CHEM 2009 presented in the SIGIR conference [16], in a paragraph entitled “Superfluous citations”, the authors report on the USPTO applicants’ behaviour, which tends to “overcite”, listing hundreds of patents as related work. Indeed, a further analysis of the query set (Table 2) shows that the USPTO patents contained an average of 66 citations to find, against 35 for WIPO patents and 25 for EPO patents. Moreover, let define *inter-connection* as, for a given patent, the percentage of its cited patents that are linked by a direct citation. For USPTO, for a given patent, inter-connection was 70%, as 70% of its cited patents were linked. Inter-correlation was 59% for WIPO patents and 57% for EPO patents, while it was only 8% for EPO patents belonging to the CLEF-IP 2010 topic set.

These facts consolidate both assumptions: we observed this over-citing behaviour in USPTO patents; but we indeed observed a strong inter-correlation in all these cited chemical patents compared to general patents at CLEF-IP. Our citations feedback strategy achieves to strongly capture the behaviour of chemical applicants, which tends to cite regular patterns of multiple patents massively inter-correlated with direct citations.

4) Filtering based on IPC codes

Once again, strategies based on USPTO IPC codes were disappointing, while such strategies were effective in the past CLEF-IP tracks [8,9] (Table 3).

IPC level	MAP at TREC-CHEM	MAP improvement at CLEF-IP
Baseline	0.261	
IPC subclasses filtering	0.222 (-15%)	+8%
IPC subgroups filtering	0.188 (-28%)	+17%

Table 3. Results for IPC filtering for both levels, compared to improvements observed at CLEF-IP.

We can make two assumptions: this strategy fails in TREC because it is applied on a specific domain (chemistry); or this strategy fails because the quality of IPC annotations in the collection (70% is from USPTO) is insufficient. However, several papers from searchers working on automatic IPC categorization, or from the WIPO itself, report on the insufficient quality of IPC annotation at the USPTO [17,18,19]. This fact consolidates the second assumption, but further experiments or expertise are needed for clarifying this issue.

5) Filtering based on Publication Dates

Finally, this simple strategy reached the performance from 0.261 to 0.266 for MAP (+2%). This was the final run that was officially submitted.

6) Query Expansion using chemical annotation for the TS task

The Average Precision (AP) of the baseline run was 0.011. Such a low global performance can be explained by the size of the collection, as experts judged only 1% of the documents submitted by all systems being relevant. However, our three different strategies can still be compared and evaluated regarding to our baseline run. The small QE run had an AP of 0.015; this result confirms the benefit of normalization on search effectiveness for drug-related information request. The benefit of large synonyms expansion is less obvious: the medium QE run had an AP of 0.011, while the large QE run finally led the best AP 0.022, which is a +100% improvement. Results analysis shows that such automatic large synonyms expansion boosted the performances for some queries (e.g. amoxicillin synonyms such as *clamoxy* or *amoxil*, see Figure 3), but also strongly degrades them for other queries. Such an approach showed promise, but needs to be strengthened by a binary classifier in order to determine if the Query Expansion is beneficial for a given query. In other scenarios, synonyms need to be manually validated in a semi-automatic pipeline.



Figure 3. Example of normalization and expansion for the official TREC 2010 query (TS-30) containing chemical named entities such as amoxicillin as recognized by the ChemTagger. Each normalized chemical term is associated with a confidence score depending of its context, a unique identifier (here from MeSH or PubChem) and preferred form with a set of synonyms (e.g. amoxil, clamoxyl...).

7) Query Expansion using IPC automatic categorizer for the TS task

Runs computed with IPC Query Expansion were unfortunately not evaluated by the assessors.

Conclusion

The 2010 edition of the TREC-CHEM track confirmed for our system the power of citations feedback for chemical Prior Art Search. This strategy achieved again significant improvements for the performances of our system at the PA task (from 0.043 to 0.261 for MAP, +507%) and started to be used by other participants. Furthermore, the longer the queries were, the more powerful the method was. Thus, the citations feedback could be quite efficient for capturing the applicant's behavior during the Prior Art at the USPTO. On the other hand, IR based on traditional approaches achieved relatively low results and seemed limited on this task, especially for these long queries. Finally, IPC codes once again were of no use, and further experiments need to determine if their quality in USPTO patents is sufficient in order to be exploited for retrieval tasks.

One limitation of the citations feedback is that it was applied on queries which were contemporary to the collection. Thus, it can exploit the references of patents

that were unknown when the application was filed. In TREC-CHEM 2009, we tried to restrict this strategy to the patents filed before the application, and this led to a -15% performance. However, it was said at the TREC Chemical Workshop that such a retrospective Prior Art Search can be seen as a useful scenario.

People involved in chemical retrieval for USPTO patents should definitely consider strategies beyond the standard Information Retrieval.

Acknowledgments

The study reported in this paper has been partially supported by the European Commission Seventh Framework Program (DebugIT project grant no. FP7-ICT 217139).

References

- [1] M Lupu, F Piroi, X Huang, J Zhu, J Tait, "Overview of the TREC 2009 Chemical IR Track", TREC 2009
- [2] TREC-CHEM 2010 Track Guidelines
- [3] <http://eagl.unige.ch/bitem/>
- [4] G Roda, J Tait, F Piroi, V Zenz, "CLEF-IP 2009: retrieval experiments in the Intellectual Property domain", CLEF 2009 Working Notes

- [5] F Piroi, J Tait, "CLEF-IP 2010: retrieval experiments in the Intellectual Property domain", CLEF 2010 Working Notes
- [6] <http://www.ir-facility.org/events/irf-symposium/2010/patolympics>
- [7] J Gobeill, D Teodoro, E Pasche, P Ruch, "Report on the TREC 2009 Experiments: Chemical IR Track", TREC 2009 Proceedings
- [8] J Gobeill, E Pasche, D Teodoro and P Ruch, "Simple Pre and Post Processing Strategies for Patent Searching in CLEF Intellectual Property Track 2009", Lecture Notes in Computer Science, vol 6241, pp 444-451, 2010
- [9] D Teodoro, J Gobeill, E Pasche, D Vishnyakova, P Ruch, and C Lovis, "Automatic Prior Art Searching and Patent Encoding at CLEF-IP '10", CLEF 2010 Working Notes
- [10] I Ounis, C Lioma, C Macdonald and V Plachouras, "Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web", Novatica/UPGRADE Special Issue on Next Generation Web Search, vol 8, pp 49-56, 2007
- [11] http://ir.dcs.gla.ac.uk/terrier/doc/configure_retrieval.html
- [12] M Porter, "An algorithm for suffix stripping", Program, vol 14, pp 130-137, 1980
- [13] P Corbett and P Murray-Rust, "High-Throughput Identification of Chemistry in Life Science Texts", CompLife 2006, LNBI 4216, pp. 107-118, 2006
- [14] <http://eagl.unige.ch/EAGL/>
- [15] Y Wang, J Xiao, T Suzek, J Zhang, J Wang and S Bryant, "PubChem: a public information system for analyzing bio-activities of small molecules", Nucleic Acids Res, 2009.
- [16] M Lupu, J Huang, J Zhu and J Tait, "TREC-CHEM: Large Scale Chemical Retrieval Evaluation at TREC", ACM SIGIR Forum, vol 63, pp 63-70, 2009
- [17] WIPO Advanced Seminar on the International Patent Classification, Newport, UK, Dec 7-11, 1998
- [18] T Xiao, F Cao, T Li, G Song, K Zhou, J Zhu and H Wang, "KNN and Re-ranking Models for English Patent Mining at NTCIR-7", Proceedings of NTCIR-7 Workshop Meeting, Dec 16-19, Tokyo, 2008
- [19] D Teodoro, E Pasche, D Vishnyakova, C Lovis, J Gobeill and P Ruch, "Automatic IPC encoding and novelty tracking for effective patent mining", Proceedings of NTCIR-8 Workshop Meeting, Jun 15-18, Tokyo, 2010
- [20] <http://pingu.unige.ch:8080/IPCCat>
- [21] <http://eagl.unige.ch/ChemTagger>