TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC

Mihai Lupu
Information Retrieval Facility
m.lupu@ir-facility.org

Jimmy Huang
York University
jhuang@yorku.ca

Jianhan Zhu
University College London
j.zhu@cs.ucl.ac.uk

John Tait
Information Retrieval Facility
john.tait@ir-facility.org

Abstract

Over the past decades, significant progress has been made in Information Retrieval (IR), ranging from efficiency and scalability to theoretical modeling and evaluation. However, many grand challenges remain. Recently, more and more attention has been paid to the research in domain specific IR applications, as evidenced by the organization of Genomics and Legal tracks in the Text REtrieval Conference (TREC). Now it is the right time to carry out large scale evaluations on chemical datasets in order to promote the research in chemical IR in general and chemical Patent IR in particular. Accordingly, we organize a chemical IR track in TREC (TREC-CHEM) in order to address the challenges in chemical and patent IR. This paper describes these challenges and the accomplishments of the first year and opens up the discussions for the next year.

1 Introduction

Any evaluation campaign has a set of criteria that generally fall in one of two categories: effectiveness (does the system do what it was designed to be doing?) and efficiency (how fast/reliable/cheap is it?). While in principle these two categories do not conflict, in practice, because human experts have to be involved in the effectiveness category, it is hard to run one experiment that goes both sufficiently deep in the analysis to assess actual effectiveness in real user context and sufficiently large scale to give a clear image of the scalability of the different systems. This is why we divided our track into two sub-tasks.

Technical Survey Task: 18 topics have been kindly provided by chemical patent experts based on their information needs. Participants' systems retrieve a ranked list of documents in response to each topic. In order to alleviate the evaluation work for the experts, and compare ordinary users and experts' views on relevant judgments, we carried a two step evaluation procedure, where each topic is judged by two graduate students majored in chemistry in the first step, then presented to a patent expert for judgments by taking into account the

students' judgments in the second step. This task enables us to understand the pros and cons of the participating systems in finding relevant chemical documents and how effectiveness can be improved.

Prior Art Search Task: The second task asks participating systems to find relevant patents with respect to a set of 1,000 existing patents. The results returned by the systems are not manually evaluated, but are assessed based on existing citations of the 1,000 patents and their family members. This task also contains a mini-task, where the participants are invited to submit the results to only the first 100 patents in the list, if their computing resources do not allow them to retrieve results for the full list of 1,000 topics by the result submission deadline. This task helps us investigate how to design both effective and efficient systems that can retrieve high quality relevant documents for a rather large number of topics.

The track organizers received registrations from 14 research groups from both academia and industry, who were allowed to download the data and topics. Eventually, 8 groups submitted at least one run to at least one of the two tasks.

The methods applied vary substantially, from basic IR methods (e.g. vector space models without any pre-processing of the text) to advanced chemistry-specific methods using named entity recognition software and synonyms of chemical substances. It will be very interesting to look into the results of these methods as soon as the experts will have completed their evaluation on the retrieval results. Until then, partial results based on student evaluation show that about 45% of the documents presented for evaluation have been judged relevant by the students. However, there is a large variance across topics. For example, only 6% documents are judged as relevant for one particular topic, while 93% documents are judged as relevant for anther topic. Even more, the two students who evaluated each topic did not always agree. In fact, almost 1 in 5 evaluations had conflicting results (not relevant versus relevant or highly relevant). The experts, who are working on these topics now, will have the final say on these controversial results, while it is interesting to understand where the disagreements arise and what can be done for better evaluations in the future.

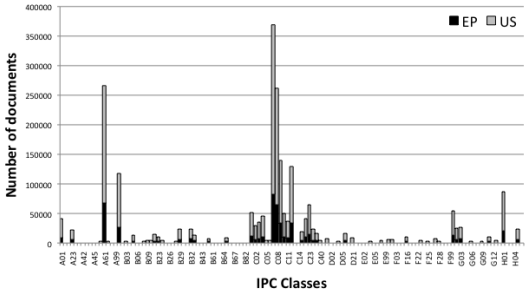
In the remainder of the paper, we give a description of the test collection in our evaluation in Section 2, present our two evaluation tasks in Section 3 and 4, respectively, propose future directions in Section 5 and conclude in Section 6.

2 Test Collection

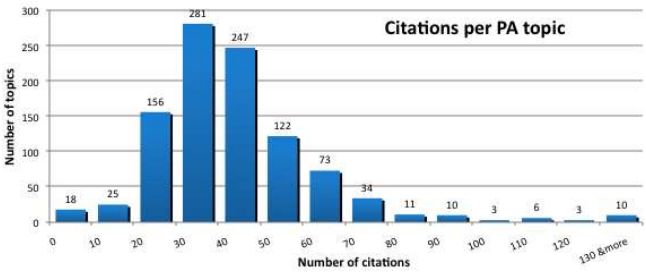
In 2009, the Information Retrieval Facility (IRF) disposed of 2,648,160 patent files (approximately 112GB in size) from the chemical domain (classified under IPC codes C and A61K) that, after pre-processing, were available to the participants. This collection covers patents in the field until 2007, registered at three major patent offices, i.e., EPO, USPTO and WIPO. The format of the patent files is in XML.

Among these 2.6 million patent files, we distributed 1,185,012 files that contain (claims and (description or abstract)) - i.e. that have enough textual information for making them more useful in text retrieval. They sum up to 98.22GB of uncompressed data. Among these files, Figure 1a shows their distribution per IPC class. Two observations are: 1. Class C is "Chemistry" 2. Many patent files are classified under more than one class, hence the apparently larger number of files in Figure 1a than the 1.185 million just mentioned.

This year's track has also benefited from chemical scientific articles, kindly provided by the Royal Society of Chemistry (RSC) in the UK. This data set consists of about 59,000



(a) Distribution of files per patent class



(b) Distribution of the number of citations per topic in the prior art search topic set

Figure 1: Some statistics on the data collection

scientific articles, for a total of size of around 3GB, from 31 journals published by the RSC. The format is also in XML, but different schemas have been used for the RSC and patent documents.

3 Prior Art Search Task

The test topics used in the task are 1,000 patent files. This is similar to real world prior art search faced by patent applicants or patent lawyers, who need to create a list of patents as prior art upon drafting or receiving a patent application, respectively. The participants are asked to retrieve a ranked list of patents as prior art for each topic. The aim is to see how well systems work on behalf of the humans in prior art search.

3.1 Evaluation

The query relevance judgements (qrels) for this task were generated based on citations not only in the patent that constitutes the topic (i.e. the hypothetical application for which we want to find prior art) but also from family members of both the cited patents and the topic patent. The distribution of the number of citations per topic is shown in Fig. (1b).

3.1.1 Citations and families

A few more details about citations and patent families are in order here. Like a research paper, a patent is, at different points in time, associated with other patents (or research papers). Before applying for a patent, the applicant must do a prior art search (as a scientist would include a Related work section in a paper submitted to a conference or journal). This list of references are called *Applicant's citations*. Then, upon receiving the application, the patent office will review it and add another set of references, possibly rejecting the patent application. The similarity with academia is again striking: upon receiving a submission, a conference's programme committee would review the paper, suggesting other related works and possibly rejecting the submission for not being sufficiently novel. This set of references, that the patent office adds, are called *Examiner's citations* or *Search report citations*. Finally, in some cases, after a patent is published, a third party (a competing company, for instance) will oppose it, referencing works that neither the applicant, nor the examiner found, but which are very related to the new patent. This is less common in academia, but you can think of it as the situation in which, while presenting the work at the conference, a member

of the audience stands up and claims that he had already solved the same problem years ago (hopefully citing some concrete works to prove that). Such references are called *Opposition citations*.

Putting all those references together, our collection of patents can take advantage of the large amount of manual work already done by field experts.

Furthermore, we can extend this by looking at patent families. In our collection, we consider "simple" families: patents are related if they are basically the same idea submitted to different patent offices for protection in different jurisdictions. It is a fact that when submitted to a patent office, there is a bias of the examiner towards the collection of that patent office, and other documents, published elsewhere, are not cited. We compensate this by considering a patent relevant in one of the following three situations:

1. it is directly cited by the topic patent,
2. it is a family member of a patent directly cited by the topic patent,
3. it is directly cited by a family member of the topic patent.

3.1.2 Pitfalls

The set of patent citations seems like a wonderful resource that we can take advantage of and in most aspects it is rightfully so. However, a number of possible pitfalls need to be taken into account when using this resource for automatic IR evaluation. We present them in this section, opening up the discussion to members of the community on the best way to tackle them.

Incompleteness. While theoretically the set of citations that constitute the Prior Art of a patent cover *the entire human knowledge, in any form and language it may have been expressed*, it is clearly not the case. What makes it more complicated in this case, is that we don't know, for each individual patent in our collection, if an expert ever looked at it and judged it not worthy to be included in the list of prior art citations or never saw it: i.e. there are no "non-relevant" judgements, only "relevant" ones.

Superfluous citations. In some cases (rather numerous in the US applications), the applicant tends to over-cite, listing hundreds of other patents as related work, while in reality they might at most have the same field of work in common. For instance, if we take the top 1000 most citing patents registered with the USPTO, we have an average of 206 citations per patent. The top 1000 most citing patents at the EPO have on average only 13.6 citations! That is more than a tenfold difference, and it is hardly arguable that US patents are less novel that EP patents. It is clearly a feature of the procedure that needs to be accounted for.

Graded relevance. From the way that these citations are generated, we can deduce some graded relevance judgements. We may consider for instance that opposition citations are highly relevant, because no one would risk opposing a patent without a strong case. We might also consider that the applicant's citations are at best relevant, since whether intentionally or non-intentionally, an applicant would not list a reference that invalidates its own patent. However, these lines are blurred and they differ from patent office to patent

Table 1: Categories of 18 technical survey task topics.

Topic area	Number of topics
pharmaceuticals	6
organic, high molecular weight	4
inorganic	2
formulations	2
emulsions	2
organic, low molecular weight	1
reaction conditions	1

office, as we have seen above. Furthermore, when using family citations, how relevant should these indirect citations be, since family members are not always quite identical?

3.1.3 Metrics

As we all know, there is a plethora of metrics that try to cover different aspects of IR evaluation and there is little agreement on even the most common ones, like recall [3] and everything derived from it (including MAP). However, the case of patent search is arguably different than general search and, since this task uses a limited collection of patent only, one could reasonably assume that the (at least) two experts that did the search (applicant and patent examiner) did in fact an exhaustive search over the collection and everything that is not cited is indeed non-relevant. Consequently, for this special case, every measure that works in the original Cranfield paradigm can be used here.

4 Technical Survey Task

Arguably, the prior art search task does not give us enough depth of analysis into specific issues of chemical information retrieval. An in-depth, albeit automatic, analysis would be hard to do on a large collection and would potentially discourage new participants from taking part in our track. We are particularly concerned in identifying technology that has potential but has not yet been optimized to operate at large scale.

The 18 technical survey topics used this year were generously provided by 5 patent experts, from their previous experience in their respective specific fields of interest. Table 1 shows the chemical categories of the 18 topics[1].

4.1 Stratified sampling

To cope with the large number of documents to be evaluated, versus the small number of available experts in the field, we had to sample out a subset of documents to be evaluated. We used a stratified sampling approach, whereby the results were pooled together and the first 10 documents were always sampled, the next 20 documents were sampled with a probability of 30% and the next 70 documents (i.e. up to 100) sampled with a probability of 10%.

[1]Some topics fall under multiple categories.

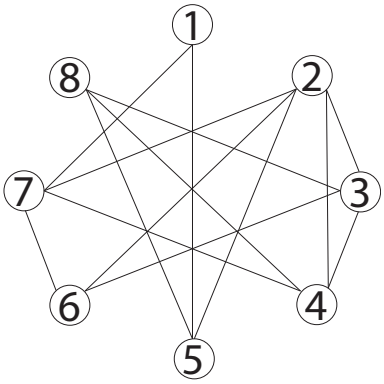


Figure 2: The Correlation of Evaluators in Student Evaluation

4.2 Student evaluation

Eight students who are majoring in chemistry and its related subjects are invited to participate in the Technology Survey (TS) task evaluation of TREC-CHEM this year. Most of these students are Ph.D. students from universities in Canada and the United States.

Each topic was evaluated by two different student evaluators. Therefore, 36 topics have been evaluated by students in total. Based on the topic categories described in Table 1, they were assigned to students mainly according to their specialized fields and interests.

In Figure 2, the 8 nodes stand for 8 different student evaluators. If the evaluators have one or more topics in common to be evaluated, then they are connected by a line. As we can see from Figure 2, each evaluator has been connected to at least two other student evaluators. This will allow us to better study inter-assessor agreement.

An online evaluation system has been used, such that the evaluators could access the documents whenever they have a computer connected to the Internet. We adapted the system developed by NIST and used in previous years by the Legal track. The relevance judgements made by the student evaluators for all the documents were graded as "non-relevant", "relevant" and "highly relevant".

From our initial analysis of the results of the TS topics, it is clear that some were much more difficult, or ambiguous, than others. As mentioned earlier, up to 93% percent of the results were judged relevant by at least one of the two students in one case (topic TS-9). We are discussing with the expert that introduced that topic to try to understand what happened, because he did not expect so many relevant results. This experience will allow us to better formulate the requirements for next year's topics list.

4.3 Expert evaluation

We presented the experts with a "lenient" merge of the evaluations of the students (i.e. if at least one student rated a document as relevant, the merged list rated it as relevant as well). Using the same graphical web interface as the students, the experts had the option to look only at those documents that the students did not judge, only at those that were judged relevant, or at the entire list. Unfortunately, the interface did not log this choice, but we rely on their comments and observations after the evaluation exercise.

Overall, the experts tended to be more strict in their interpretation of relevant, and reduced the number of relevant documents. In a couple of cases, this reduction was extremely drastic (from a couple of hundreds to 32 in one case, or from 160 to only 4 in another case).

This is source for though for next year's track, as we have to understand whether the topic was under-specified, or the student evaluators were not knowledgeable enough.

5 Future directions

As for any new track, there will be much to be improved in the second year. One of our strategic decisions at the beginning of last year was to keep it basic in terms of the evaluation procedures and measures, collection and requirements on the participants. We have required only a basic ad-hoc search, evaluated using trec_eval 9.0 (PA task) and inferred MAP and nDCG (TS task) [2], we did not require participants to use a specific ontology or entity recognition system. Nevertheless, the manual topics were varied and contained a number of chemistry-specific issues, such as

- pattern structures (markush structures)
- numeric ranges
- roles of chemicals (e.g. documents containing some chemical as an anti-cancer agent, not anti-cancer pain)
- reactions

Based on the experts feedback, still to be received by the organizers, we will be able to decide which areas among these are of further interest. At the IRF, we have taken a keen interest in this and our October newsletter issue [1] was dedicated to chemistry information retrieval.

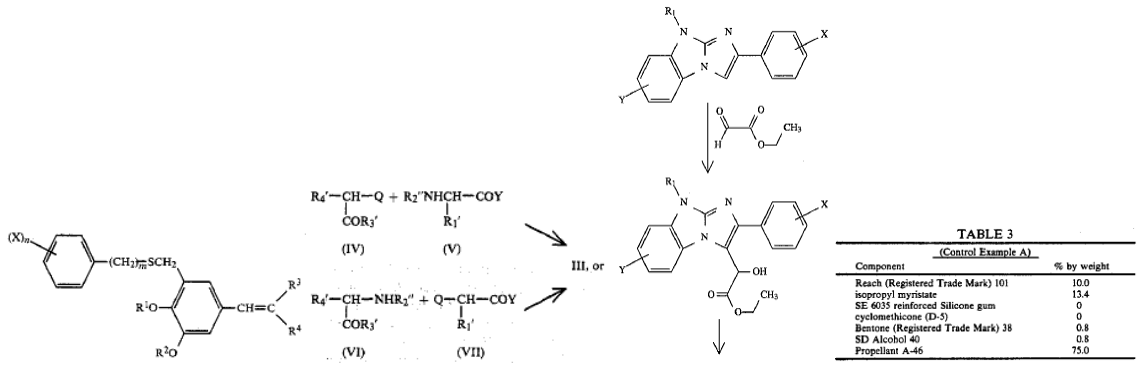
Each of the chemistry-specific issues above were identified by experts of being in the set of "unfulfilled wishes", and this year's experience adds to our understanding, but does not solve them. More experiments, more topics and more evaluation is needed to decide on reliable measures to assess the quality of chemical IR systems.

Apart from these, the TREC-CHEM organizers are considering a number of new issues, to be introduced in the second year:

- *Image recognition.* We have the capacity in the second year to add images to our collection of documents. These images may contain chemical formulas (Figure 3a), chemical reactions (Figure 3b), chemical reaction flows, possibly spanning several pages (Figure 3c) or tables containing chemical parameters (Figure 3d).
- *Entity retrieval.* Rather than retrieving documents, participants will be asked to return a specific compound or list of compounds.
- *Passage retrieval.* Rather than retrieving full documents, ask the participants to highlight the salient passages in the document.
- *Interactive retrieval.* A real searcher, when doing a technology survey or prior art search, does not issue one query and expect to get all results by that one interaction with the system. Having several rounds of interaction helps, but the experts must dedicate even more time to interact with the research groups.

6 Conclusion

In this paper we presented an overview of the first year of the TREC Chemistry track, with its problems, achievements and issues for the next year.



(a) Formula (b) Reactions (c) Process flows (d) Tables

Figure 3: Some images from chemistry patents

The main challenge we face is the lack of domain experts to evaluate the participants' results. Using students to do a prior filtering of clearly non relevant documents yield promising results, but also showed that there are, in some cases, distinct differences between judgements done by experts and those done by students. This issue still requires investigation.

We try to strike an optimal balance between the needs of the experts and the capacities of the current IR systems. Avoiding raising hopes beyond what is achievable, and thus disappointing and losing end users is one of our concerns, while raising the entry bar for participants just high enough to allow fresh ideas from research groups who may not have worked in this domain before, is the other one. This year's Chemistry Track has been a success thanks to many people and we invite all interested researchers in contacting the organizers with ideas, suggestions or just expressions of interest for the next year.

Acknowledgements

We wish to thank Ellen Voorhees and Ian Soboroff at NIST for their numerous advices and for providing the web tool used for evaluation; Michael Siu at Vice-President Research Office of York University for his support and advice; Emine Yilmaz at Microsoft Research and Evangelos Kanoulas at Northeastern University for their help with stratified sampling and evaluation of the TS tasks; and last but not least, Florina Piroi at the IRF who helped us tremendously in processing and handling all the results and evaluation files.

References

[1] Information Retrieval Facility Newsletter - Special Issue: Chemistry. http://www.ir-facility.org/the_irf/newsletter, October 2009.

[2] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2008. ACM.

[3] J. Zobel, A. Moffat, and L. Park. Against Recall: Is it Persistence, Cardinality, Density, Coverage, or Totality. *SIGIR Forum*, 43(1), June 2009.