

# TREC-CHEM 2010 : Notebook report

Mihai Lupu, John Tait, Jimmy Huang, Jianhan Zhu

October 24, 2010

## Abstract

The TREC Chemical IR Track is a domain-specific evaluation campaign working with documents containing specific lexica, including chemical formulas and specific names. The 2010 edition of the track also included supporting material in addition to text: images and structure information files. As in the previous year, we had two tasks: a patent focused prior-art (PA) task and a user-focused Technology Survey task (TS). The data collection includes patent files as well as scientific articles, together with their attachments, if any. Topics and relevance judgments were either automatically or manually created.

## 1 Introduction

The 2nd TREC Chemical IR track follows closely on the principles and objectives outlined in the first edition. Through two tasks, it aims to cover both the issues of large scale retrieval, as well as in-depth analysis of the chemical domain. Like last year, one task (*Prior Art*) asked the systems to find relevant patents with respect to a set of 1,000 existing patents. The results returned by the systems were evaluated based on the citations in those patents and their family members. As last year, we selected a subset of 100 patents to be the *Small PA* task, for those systems which could not provide answers to the full set.

The second task, the *Technology Survey*, was designed to mimic in closer detail the kinds of queries issued by experts in the field. This year, 5 [patent] experts kindly provided a total of 30 such topics. Participants were then asked to provide results from the full data set, which this year included almost 200,000 scientific articles. The results are then evaluated manually.

Similarly to the last year, in 2010 we provided the data collection to a total of 13 groups, mostly from academia but also from industry. However, we received much fewer results back: 4 groups submitted 11 runs for the PA task, while only 2 submitted 12 runs results for the TS task. Overall, the methods applied are also similar to last year: entity recognition, standard BM25, re-ranking methods based on citations.

The remainder of the paper is organized as follows: Section 2 describes the data collection, then Sections 3 and 4 provide details on the PA and the TS tasks, respectively. A summary of approaches, as provided by the

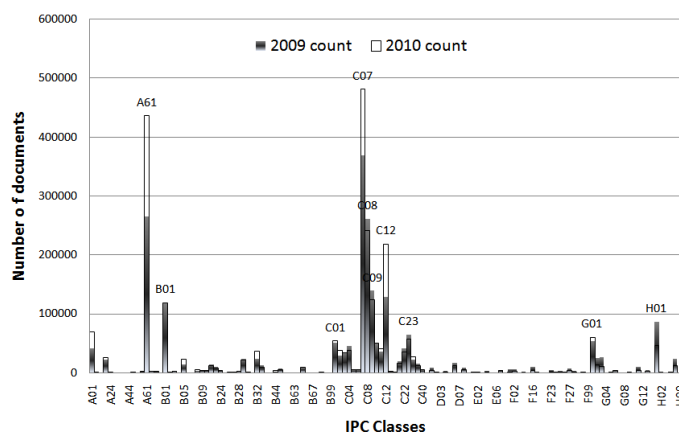


Figure 1: Distribution of patent documents over IPC classes

Table 1: Distribution of patent documents across sources

Source	Number of documents
EPO	134035
USPTO	907170
WIPO	236262

participants is listed in Section 5 and the conclusion and outlook are in Section 6.

## 2 Data Collection

The TREC-CHEM10 collection contains 1,277,467 patent documents and 176,528 scientific articles, which, together with their attachments (images, chemical structure files, pdfs) totaled 420GB of compressed data, that we made available for download. The download site was structured on type of file, so groups which knew that they wouldn't be able to process certain types, only downloaded the files they could actually work with. To the best of our knowledge at this point, none of the groups worked with anything other than text data in xml files.

For the patent files, the distribution across different chemical domains is shown in Figure 1, for both the 2009 and 2010 collections. As can be seen, the emphasis on particular classes is maintained, with an increased emphasis on the biomedical field (class A61K). For the scientific articles collection, in addition to the articles we had in the 2009 collection from the Royal Society of Chemistry (31 journals), we added a lot more articles from PubMed Central Open Access collection and from four individual publishers (Hindawi, Oxford, IUCrJnls, MDPI), totaling over 1500 journals. However, given the new sources of data, we also observe in this part of the collection, an emphasis on the bio-medical part of chemistry,

to the detriment of other sub-domains. Still, this can be explained by the fact that there is indeed a significantly larger industry in real life on this kind of chemistry.

**Unique identifiers.** Each document must have a unique identifier. For the patent domain, this is the UCID which consists of the identifier of the issuing office, a number, and a version identifier (also called a *kind* code). The kind code, to put it simply, identifies the different versions of a patent document, as it evolves in the granting procedure. Unlike last year, where we considered a patent to be any version of the document, basically using only the country code and the document identifier, this year we used the full UCID.

For scientific articles, the unique identifier is the Digital Object Identifier (DOI).

### 3 Prior Art (PA) Task

One of the lessons we learned last year with respect to the design of the PA task, was that, given the way we evaluated (using mostly citations from the examiner’s report on the novelty of the patent application), we must provide as topics the application documents rather than the granted patents. We also had learned that our sampling procedure had been biased towards the US patents, so this year we corrected both of these issues and resulted in an equal distribution over the 3 sources of patents used in this collection: 333 from the USPTO, 333 from the WIPO and 334 from the EPO. The small topic set is slightly more imbalanced, as we selected it at random from the large set, without imposing specific limits with regards to the source: 27 from EPO, 36 from the USPTO and 37 from WIPO.

The topics for this PA task are represented by the full text of the patent application documents. The text provided to the participants contains everything, with the exception of the legal status of the document, since this is not part of the research collection. However, the legal status is irrelevant for this track. However, the text did contain the citations that the applicant or the patent examiner had added. Participants were instructed not to use them directly in their ranking process. However, given that the citations in the collection provide useful information, we did not limit the participants use of other references in other documents.

#### 3.1 Relevance judgements

Like in 2009, the qrels for the PA task are created based on citations within the patent document, citations provided by applicant, patent office, or during an opposition procedure. The procedures is done in three steps, as follows:

- D: contains the citations recorded in the topic patent itself. These are also called ‘direct citations’;

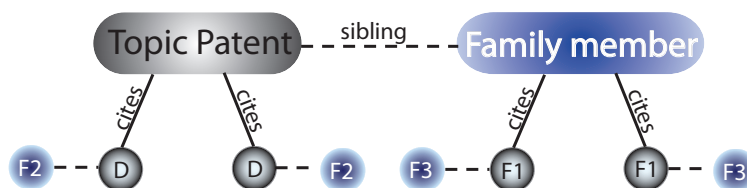


Figure 2: Qrels creation procedure

F1: contains the direct citations, D, to which we add the citations recorded in the family members<sup>1</sup> of the topic patent;

F2: contains the direct citations, D, to which we add the family members of the documents in the D set;

F3: contains the F1 and F2 sets to which we add the family members of the patent documents in the F1 set.

Figure 2 illustrates this process, with the observation that it only marks as F1, those patents that appear after the first step, with F2 after the second and with F3 after the final step. In reality, there is a relation of inclusion between the four sets:  $D \subset F1 \subset F3$  and  $D \subset F2 \subset F3$ . The F3 was used as the final QREL set, after having filtered out the documents which were not in the collection.

### 3.2 Results

We calculated results for the full and Small PA topic set for 6 different measures: MAP, Reciprocal Rank, Precision at 30, recall at 100, and NDCG. The results are shown in alphabetical order of the runs in Tables 2 and 3. The tables do not include the two runs submitted by the Iowa team, as there was clearly something wrong with those runs, as none of them contained any relevant documents whatsoever, resulting in a pure zero score across the board. To avoid skewing the averages, in agreement with the team, we decided to not include them in this report. In addition, the SCAI team requested a change of their submitted runs after the deadline, as they had inadvertently used the direct citations of the topic patents in their results. In agreement with the NIST organizers, we accepted this change and the results provided here reflect these new, lower scores compared with their original submission results.

For the full PA set, we also computed the statistical significance of the difference between MAP and NDCG results of the submitted runs, using the randomization test. The results, shown in Tables 4 and 5 show the p-values and indicate that all runs are significantly different, with the exception of the (SCAI10NRMTOK, York10CaPA01) pair in the case of the NDCG.

---

<sup>1</sup>A 'patent family' is a set of patents granted by different patent authorities but related to the same invention.

Table 2: Results on 6 measures for the full set of PA topics

measure	map	bpref	recip rank	P 30	recall 100	ndcg
BiTeM10PAx	0.2657	0.6592	0.6121	0.3485	0.4724	0.4975
SCAI10CIENTP.result	0.4121	0.7075	0.7153	0.4554	0.5491	0.5834
SCAI10CITENT.result	0.2336	0.5468	0.5324	0.2794	0.3596	0.4119
SCAI10CITNP.result	0.2065	0.5110	0.4769	0.2485	0.3265	0.3764
SCAI10CITTOK.result	0.0947	0.2804	0.1956	0.1126	0.1511	0.1888
SCAI10NRMENT.result	0.0665	0.4171	0.3456	0.1169	0.1974	0.2547
SCAI10NRMNP.result	0.0551	0.3702	0.3088	0.1000	0.1743	0.2224
SCAI10NRMTOK.result	0.0172	0.1536	0.1133	0.0366	0.0629	0.0868
York10CaPA01	0.0136	0.1681	0.1022	0.0309	0.0583	0.0885

Table 3: Results on 6 measures for the Small set of PA topics

measure	map	bpref	recip rank	P 30	recall 100	ndcg
BiTeM10PAx	0.2175	0.6647	0.5102	0.2567	0.4543	0.4295
BiTeM10PAx	0.2174	0.6647	0.5027	0.2567	0.4560	0.4285
SCAI10CIENTP.result	0.3612	0.7063	0.6452	0.3617	0.5456	0.5193
SCAI10CITENT.result	0.1878	0.5157	0.4442	0.2137	0.3368	0.3450
SCAI10CITNP.result	0.1685	0.4936	0.4480	0.1960	0.3073	0.3262
SCAI10CITTOK.result	0.0554	0.2222	0.1274	0.0633	0.1059	0.1257
SCAI10NRMENT.result	0.0750	0.4718	0.3421	0.0943	0.2651	0.2502
SCAI10NRMNP.result	0.0648	0.4434	0.3244	0.0797	0.2406	0.2300
SCAI10NRMTOK.result	0.0150	0.1660	0.0730	0.0197	0.0708	0.0726
York10CaPA01	0.0132	0.1449	0.0862	0.0223	0.0550	0.0646

Table 4: p-values for the randomization test on the MAP values

run name	BiTeM10PAx	SCAI10CIENTP	SCAI10CITENT	SCAI10CITNP	SCAI10CITTOK	SCAI10NRMENT	SCAI10NRMNP	SCAI10NRMTOK	York10CaPA01
BiTeM10PAx	x	0	0.00037	0	0	0	0	0	0
SCAI10CIENTP		x	0	0	0	0	0	0	0
SCAI10CITENT			x	0	0	0	0	0	0
SCAI10CITNP				x	0	0	0	0	0
SCAI10CITTOK					x	0.00011	0	0	0
SCAI10NRMENT						x	0	0	0
SCAI10NRMNP							x	0	0
SCAI10NRMTOK								x	0.01808

Table 5: p-values for the randomization test on the NDCG values

run name	BiTeM10PAx	SCAI10CIENTP	SCAI10CITENT	SCAI10CITNP	SCAI10CITTOK	SCAI10NRMENT	SCAI10NRMNP	SCAI10NRMTOK	York10CaPA01
BiTeM10PAx	x	0	0	0	0	0	0	0	0
SCAI10CIENTP		x	0	0	0	0	0	0	0
SCAI10CITENT			x	0	0	0	0	0	0
SCAI10CITNP				x	0	0	0	0	0
SCAI10CITTOK					x	0	0.00027	0	0
SCAI10NRMENT						x	0	0	0
SCAI10NRMNP							x	0	0
SCAI10NRMTOK								x	0.6976

## 4 Technical Survey (TS) Task

The TS task is similar to a traditional ad hoc retrieval task, however, the challenge is the way to deal with chemical specific problems such as synonyms and abbreviations. Five patent and academic chemical experts have kindly provided 30 topics from their experience.

As mentioned in Section 2, the 2010 collection contained significantly more data than the 2009 one. This may be one of the causes for which we only received runs from 2 groups. Although together they summed 12 runs, the fact that all of these came from only two groups lead us to the decision not to evaluate the full set of topics, as the results, obtained through the usual pooling technique, would have been skewed towards these two groups and thus potentially less useful for future evaluations. However, given that these two groups did a considerable amount of work, we decided to provide relevance judgements to 6 of these topics, in order for them to further improve their systems.

Unfortunately, at the time of writing of this report, the evaluation results for these 6 topics are not yet available.

Unlike the previous year, when we asked students to evaluate results independently from the experts, and then passed these results to the experts, this year we designed a new interface that allows a junior evaluator to communicate directly with the creator of the topic, considered the expert on the matter, and quickly clarify any problems. This was a result of the observation last year, that 1. the experts often did not trust the students' evaluation and re-did everything and 2. even a limited interaction between the junior evaluator and the expert resulted in a significant agreement between the two in the final results. The new evaluation interface is briefly described in what follows.

### 4.1 Evaluation interface

The two main differences between the 2010 and the 2009 assessment interface are:

1. a discussion forum (Figure 3)
2. a "notes" section (Figure 4)

Both of the new features work on a topic level and are shared between the two evaluators of the topic (the junior evaluator and the expert). The discussion is aimed to provide a platform where questions and answers are communicated and recorded. Whenever one of the two uses the feature an email is sent to the other, as well as to the administrators (i.e. organizers).

This feature not only helps the junior evaluator provide a better evaluation of the topic by a better understanding of the problems addressed, but also saves for future analysis all the issues which were discussed and therefor helps the organizers make better topics in the future.

In relation to this last objective, and as a consequence of requests from evaluators last year, we added a feature which allows the user to take notes during the evaluation. Again, these notes are shared by all those who have access to the same topic.

**ChemAssess - working on topic 36**

Start assessment... Log out Help

**Title:** selective activation and application of ammonia in homogeneous catalysis  
**Narrative:** The users are looking for information about selective activation and application of ammonia in homogeneous catalysis. The formation of amino compounds from ammonia is one of the crucial biological processes in the nature to utilize the nitrogen element, yet poor effort has been devoted to activate ammonia directly in organic synthesis. It is believed that

Show:  only unjudged/unsure  only relevant  entire pool

1: [EP-0209807-A2](#)  
**Process for the preparation of high-temperature resistant compact or cellular polyurethane elastomers.**  
 unjudged  unsure  not relevant  relevant  highly relevant

2: [US-20050043213-A1](#)  
**Catalysis of the cis/trans-isomerisation of secondary amide peptide compounds**  
 unjudged  unsure  not relevant  relevant  highly relevant

3: [US-7205423-B1](#)  
**Process for the preparation of organo-molybdenum compounds**  
 unjudged  unsure  not relevant  relevant  highly relevant

4: [US-6267864-B1](#)  
**Field assisted transformation of chemical and material compositions**  
 unjudged  unsure  not relevant  relevant  highly relevant

5: [US-20070149807-A1](#)  
**PROCESS FOR HETEROGENEOUSLY CATALYZED PARTIAL GAS PHASE OXIDATION OF PROPYLENE TO ACRYLIC ACID**  
 unjudged  unsure  not relevant  relevant  highly relevant

**Discussion on topic 36**

New message:

Previous messages:

05 Oct 2010 23:20	While looking at document US-4864041-A: While looking at document US-4052392-A: I think this patent could be considered relevant since it does contain a amination reaction of cyanuric chloride using ammonia directly, although there was only a base (NaOH) used to facilitate the reaction.
05 Oct 2010 23:11	While looking at document US-4864041-A: While looking at document US-4864041-A: In this case the patent did not include any information about activating ammonia specifically using their metal catalysts, so it is fair to consider this document irrelevant.
05 Oct 2010 20:52	While looking at document US-4052392-A: The mechanism and potential activation is unclear, I think it is not relevant
05 Oct 2010 19:11	While looking at document US-4864041-A: "The organic substrates also includes amines,..." no idea if also ammonia could be activated, it is not explicitly written. So this document is potentially interesting, but as there is no explicit description in it the document is not relevant
30 Sep 2010 18:20	While looking at document WO-2007046721-A2: In this case ammonia was used as a reactant, thus the patent could be considered as relevant to the topic.
30 Sep 2010 17:44	While looking at document US-6696026-B2: In this case, the patent mainly describes production of ammonia instead of its applications. Although it contains the key words of the topic and may be consequently considered as 'relevant' by an automated IR system, hardly any correlation is

Discuss this topic

Click the button above to message the expert. If a document is selected in the pool, this will be automatically added to your message.

List chemical names Add chemical

Figure 3: Discussion on a topic using the ChemAssess Interface

ChemAssess - working on topic 21

Start assessment... Log out Help

**Title:** Uses of DHEA to improve human ovulation  
**Narrative:** Human ovulation is controlled by a number of hormones within the body. Women are interested in knowing their fertile period and in methods to improve their ovulation rate and accuracy. We are interested in the use of dehydroepiandrosterone (DHEA) to improve human ovulation.  
**Relevance:** A document will be considered RELEVANT when it describes uses or consequences of the

Journal Information  
 Journal ID (nlm-ta): *Reprod Biol Endocrinol*  
 ISSN: 1477-7827  
 Publisher: BioMed Central, London

Article Information  
 Link:  
 Copyright © 2003 Fraser and Wulff; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.  
 Received Day: 29 Month: 4 Year: 2003  
 Accepted Day: 10 Month: 11 Year: 2003  
 collection publication date: Year: 2003  
 Electronic publication date: Day: 10 Month: 11 Year: 2003  
 Volume: 1 First Page: 88 Last Page: 88  
 ID: 305342  
 Publisher Id: 1477-7827-1-88  
 PubMed Id: 14613536  
 DOI: [10.1186/1477-7827-1-88](https://doi.org/10.1186/1477-7827-1-88)

Click the button above to message the expert. If a document is selected in the pool, this will be automatically added to your message.

Show:  only unjudged/unsure  only relevant  entire pool

1: [US-4244949-A](#)  
**Manufacture of long term contraceptive implant**  
 unjudged  unsure  not relevant  relevant  highly relevant

2: [US-20030004213-A1](#)  
**Medicament against dysmenorrhoea and premenstrual syndrome**  
 unjudged  unsure  relevant  relevant  highly relevant

3: [US-20050074767-A1](#)  
**Genes associated with obesity and methods for using the same**  
 unjudged  unsure  not relevant  relevant  highly relevant

4: [WQ-1986000078-A1](#)  
**INHIBIN ISOLATED FROM OVARIAN FOLLICULAR FLUID**  
 unjudged  unsure  not relevant  relevant  highly relevant

5: [US-7205281-B2](#)  
**Process for the synchronization of ovulation for timed**

Angiogenesis in the corpus luteum  
 Hamish M Fraser<sup>1</sup> Email: [h.fraser@hru.mrc.ac.uk](mailto:h.fraser@hru.mrc.ac.uk)  
 Christine Wulff<sup>2</sup> Email: [c.wulff@onlinehome.de](mailto:c.wulff@onlinehome.de)  
<sup>1</sup>Medical Research Council Human Reproductive Sciences Unit, Centre for Reproductive Biology, 49 Little France Crescent, Edinburgh, EH16 4SB, UK  
<sup>2</sup>Department of Obstetrics and Gynaecology of the University of Ulm, Prittwitzstrasse 43, 89075 Ulm, Germany

Abstract  
 The corpus luteum (CL) is a site of intense angiogenesis. Within a short period, this is followed either by controlled regression of the microvascular tree in the non-fertile cycle, or maintenance and stabilisation of the new vasculature a conceptual cycle. The molecular regulation of these diverse aspects is examined. The CL provides a unique model system in which to study the cellular and molecular regulation of angiogenesis. Vascular

List chemical names Add chemical Hide List  
 L-arginine alpha-ketoglutarate  
 Chemical notes can be quite general and fairly long

Figure 4: Notes on a topic using the ChemAssess Interface



## 5 Approaches

### 5.1 BiTeM Group

#### For PA runs:

DOCUMENT REPRESENTATION FOR DOCUMENTS : title, abstract, claims, description and complete IPC codes were used. DOCUMENT REPRESENTATION FOR TOPICS : title, abstract, claims, description and complete IPC codes were used. Claims and description were truncated to first 10000 characters. INFORMATION RETRIEVAL : Terrier, PL2 weighting scheme (Poisson), Query Expansion POST-PROCESSING STRATEGIES : reranking with documents' citations, documents' IPC codes and documents's dates.

#### For TS runs:

Standard IR strategies, as well as different levels of chemical query expansion: small, medium and large.

### 5.2 SCAI Group

#### For PA runs:

All runs uses citation re-ranking, with or without thresholding. Semantic searches (chemical), text searches and noun-phrase searches were used in each of the two sets.

### 5.3 York Group

#### For PA runs:

Language model with Dirichlet smoothing ( $\mu = 100$ ).

#### For TS runs:

Different weighting functions (BM25, DFR, Language Model), as well as query expansion using chemical information.

## 6 Conclusion

This years TREC-CHEM track has been challenging from several perspectives. Collecting all the data was an issue at the beginning of the year, then it became an issue for the participants, as there was too much of it. We were satisfied to see how, for the PA task, those runs which used domain specific data really performed much better than those with generic approaches. We were less satisfied with the technique that does re-ranking based on citation patterns. While being perfectly valid, it is a generic technique which does not fit perfectly within the objectives of this campaign. Together with the participants, evaluators and other interested parties, we organized a workshop before the conference and shall

continue to analyze the best ways to go forward and into the third year of this effort.

We are particularly committed to make this evaluation campaign relevant to professional users and therefore will continue to push for a stronger integration between text mining techniques and structure search. All the prerequisites are now in place for a strong performance in the third year, given that we will not change the collection and re-use many of the topics of the TS tasks which were not used this year.