How to Make Manual Conjunctive Normal Form Queries Work in Patents Search

Le Zhao and Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

Abstract

This year we focused on the Technology Survey (TS) task: Given a natural language description of the topic, look for related patents about that topic. The task is close to an ad hoc retrieval task, except for the additional information of the specific chemicals or chemical reactions that the user cares about. Since there are only 6 topics for the TS task, this notebook paper is more of a case study report, than the ordinary TREC report with significance tests. We found that on average, with the infAP measure, manually created conjunctive normal form queries performed similarly as automatic keyword search with some tuning of term weights. Manual queries do not seem to always help, especially when initial keyword performance is high, but can give large improvements on difficult queries. We also used the same querying strategy in the Patent Olympics 2011 ChemAthlon task, and also include some of the ChemAthlon cases in this report. Since CNF queries are strictly more expressive than keyword queries, we try to identify problems that may have caused the manual CNF queries to be seen sometimes performing worse than the automatic keyword queries.

1 Query Formulation

1.1 Boolean CNF style queries

User information need can usually be broken down into a number of concepts that any relevant document must include in order for it to be relevant. This allows the searcher to break down the information need into a set of concepts conjoined together. Each concept would typically be represented by a number of different natural language descriptions (words or phrases or windowed occurrences) that people would use to describe these concepts in relevant documents. This results in a conjunctive normal form styled query.

This kind of query has been used widely by search professionals like lawyers or librarians, as they provide more flexibility and expressiveness in matching potentially relevant documents than the bag of word (keyword) queries. A list of references for the use of CNF queries can be found in Le Zhao's thesis proposal [3].

1.1.1 Example query

For example, in Indri query language, the query (topic TS-20) about "tests for HCG hormone especially in pregnancy tests" can be expressed as.

```
#combine(
    #syn(HCG #1(Human Chorionic Gonadotrophin) #uw2(Chorionic Gonadotropin) Choriogonadotropin Choriogonin)
    #syn(pregnancy pregnant women woman fertilization conception)
    #syn(test check detection detect)
    #syn(#syn(blood ) #syn(urine Urinalyses Urinalysis))
)
```

The #combine operator is a probabilistic AND operator, while the #syn operator treats all included terms as the same term for retrieval purposes (merging the inverted lists of the terms into one disjoined inverted list). #1 is the ordered window operator with maximum distance between words being 1 (phrase operator), #uw2 is the unordered window operator requiring all terms to appear in a text window of size 2 words.

The practitioners have known for a long time that carefully constructed CNF style queries perform better than the keyword queries. Our participation in the Chemical track Technology Survey task is to understand how high quality CNF style queries can be created, and whether they do outperform their keyword counterpart.

1.2 Keyword queries

We used simple keyword queries as baseline to compare the manual Boolean queries with.

The run CMUTStncs creates queries by aggregating all words from title, narrative and details fields. Standard stopwords were removed from the queries. For example, the generated query for topic TS-22 is:

#combine(tests for HCG hormone The hormone Human Chorionic Gonadotrophin HCG is produced when a women becomes pregnant Tests are usually carried out by analysing blood or urine We are looking for articles and patents on these pregnancy test kits or the chemical tests used to produce them Human Chionic Gonadotrophin HCG pregnancy Human Chionic Gonadotrophin HCG)

The run CMUTStncws creates queries by weighting words from title by 0.3, words from narrative by 0.6 and words from details by 0.1. The weights are trained on the TREC 2010 Chemistry track TS task. The query generated for topic TS-20 is as follows:

```
#weight(
    0.3 #combine( tests for HCG hormone )
    0.6 #combine( The hormone Human Chorionic Gonadotrophin HCG is produced when a women becomes pregnant Tests are usually carried out by analysing blood or urine We are looking for articles and patents on these pregnancy test kits or the chemical tests used to produce them )
    0.1 #combine( Human Chionic Gonadotrophin HCG pregnancy Human Chionic Gonadotrophin HCG )
)
```

2 Experiments

We used Indri search engine of the Lemur toolkit [1] to index all the patent documents. The whole interaction process of querying the index, examining top results and modifying the query was done on the Lemur CGI Web interface using Indri query language.

Retrieval model parameter mu for Dirichlet smoothing was set at 7000 based on training performance on the TREC 2010 Chemistry track TS task.

2.1 Observations

Performance statistics are listed in Table 1.

Weighted combination of words from the different query fields is consistently better than the simple keyword merging approach. Manual Boolean CNF style queries are not always better than the best keyword queries, and the average performance is very similar to the best keyword results.

Table 1. Average retrieval performance of manual Boolean vs. automatic keyword queries on the 6 TS task topics.

Boolean (CMUTSmans)		Keyword (CMUTStncws)		Keyword (CMUTStncs)	
infAP	0.1887	infAP	0.1902	infAP	0.1790
infNDCG	0.4479	infNDCG	0.4696	infNDCG	0.4421
iP10	0.6167	iP10	0.5500	iP10	0.5500
iP100	0.3250	iP100	0.3028	iP100	0.3068
iP1000	0.0866	iP1000	0.0963	iP1000	0.0868
inum_rel_ret	519.5158	inum_rel_ret	578.0244	inum_rel_ret	520.7487
inum_rel	1521.4541	inum_rel	1521.4541	inum_rel	1521.4541
num_ret	5072	num_ret	6000	num_ret	6000

3 Error Analysis

In this section, we try to identify reasons why the more expressive CNF style queries can sometimes performs worse than its corresponding keyword query. We will report performance of the different types of queries on one topic, observe the corresponding manual CNF query, analyze its performance and try to point out possible reasons that cause it to perform lower than the keyword query. We also try to fix the Boolean query in straightforward ways to facilitate our analysis. However, these observations are not generalizable conclusions, which would need to be verified against test data.

3.1 Topic TS-22

Performance for topic TS-22:

Boolean (CMUTSmans)		Keyword (CMUTStncws)		Boolean (unsubmitted)	
infAP	0.2146	infAP	0.3312	infAP	0.3607
infNDCG	0.5537	infNDCG	0.7515	infNDCG	0.6546
iP10	0.2000	iP10	0.3000	iP10	0.4000
iP100	0.0600	iP100	0.1050	iP100	0.0867
iP1000	0.0121	iP1000	0.0115	iP1000	0.0117
inum_rel_ret	12.0501	inum_rel_ret	11.5485	inum_rel_ret	11.6668
inum_rel	12.4286	inum_rel	12.4286	inum_rel	12.4286
num_ret	1000	num_ret	1000	num_ret	1000

Topic title: *Uses of hormones in detection of menopause*.

Manual Boolean CNF style query for the CMUTSmans run:

```
#weight(
    0.5 #combine(
        #uw20(
            #syn(invention method device kit)
            #syn(    menopause
                     #1(change of life)
                     #uw5( #syn(end cessation cease final last) #syn(menstrual MENSTRUATION menses) )
            )
            #syn(detection detect test check predict determine determination)
        )
        #syn( #syn( #1(Luteinizing Hormone) LH ICSH Lutropin Luteozyman Luteoziman #1(Interstitial Cell Stimulating
    Hormone))
            #syn(#1(Follicle stimulating hormone) FSH Follitropin)
        )
```

```
        )
        1.0 #combine(
                menopause
                #syn(detect detection test predict)
                #syn(   #syn(#1(Luteinizing Hormone) LH ICSH Lutropin Luteozyman Luteoziman #1(Interstitial Cell
        Stimulating Hormone))
                        #syn(#1(Follicle stimulating hormone) FSH Follitropin)
                )
        )
)
```

Analysis: This Boolean query performed worse than the corresponding keyword query even at top ranks. When creating the query, the #uw20 (occurrence in a window of size 20 words) node was first instantiated as a #band (Boolean AND) node, and results seem to include some false positives at top rank, so I restricted it down to a 20-word window. The second #combine node in the query was intended as a back-off query to match more documents. From the retrieval performance, even though the first #combine node matches 373 documents, it is still filtering out quite some relevant documents, decreasing infAP. Changing the #uw20 node back to #band (and removing the second #combine node) actually improves infAP to 0.3607. The new CNF query can match 3278 documents.

Cause: It takes a lot of time for someone unfamiliar with chemistry to examine top results. Because there are many different query formulations to try, on average, only a little bit of time could be spent on carefully examining the top results for one query. Because human are not very good at remembering the exact performance of a previous query, it could become difficult for the user to decide whether the performance is improved by a specific modification to the query, unless the change caused a huge difference in top precision. When there are many top false positives, and attempts to improve retrieval by CNF expansion is not seen to significantly improve top precision, the user may naturally want to further restrict the query. However, the results often times only get worse, because relevant results are being filtered out together with the false positives.

The interactive search interface can be improved to help the user recognize which documents are relevant more easily and which run is better more easily. Highlighting of query terms in result documents can speed up the relevance judgement process. Recording down which documents have already been judged and side-by-side rank list comparisons may make it easier for the user to compare runs.

3.2 Topic TS-33

Performance for topic TS-33:

Boolean (CMUTSmans)		Keyword (CMUTStncws)		Boolean (unsubmitted)	
infAP	0.0854	infAP	0.1175	infAP	0.1583
infNDCG	0.1634	infNDCG	0.3079	infNDCG	0.2841
iP10	1.0000	iP10	1.0000	iP10	1.0000
iP100	0.7992	iP100	0.6298	iP100	0.6325
iP1000	0.1023	iP1000	0.1724	iP1000	0.1771
inum_rel_ret	102.2907	inum_rel_ret	172.4059	inum_rel_ret	177.1481
inum_rel	678.3752	inum_rel	678.3752	inum_rel	678.3752
num_ret	1000	num_ret	1000	num_ret	1000

Topic title: *Respiratory tract disorders treatment using inhalation of porous particles containing hydrophobic amino acid and endogenous phspholipids*.

Manual Boolean CNF style query for the CMUTSmans run:

```
#combine(
    #syn(inhalation inhale)
    #syn(respiratory Pulmonary lung Bronchial bronchioles pharynx trachea alveoli alveolar)
    #syn(disorder disease infection Neoplasm Fistula Granuloma)
    #weight(
        0.5 #uw50(
            #syn(#uw5(hydrophobic #1(amino acid)) leucine isoleucine Alloisoleucine phenylalanine Endorphenyl valine Methionine Racemethionine Pedameth Liquimeth Tryptophan Levotryptophan Ardeydorm Ardeytropin Trofan Tryptacin Tryptan Optimax Lyphan Naturruhe Cysteine #1(Half Cystine) #1(Zinc Cysteinate))
            #syn(#uw5(endogenous phospholipids) phosphatidylcholines #1(Choline Phosphoglycerides) Phosphatidylcholine #1(Choline Glycerophospholipids) phosphatidylethanolamines Ethanolamineglycerophospholipids Cephalins #1(Ethanolamine Phosphoglycerides))
        )
        0.8 #uw50(
            (same two chemicals as above)
        )
    )
)
```

Analysis: The manual Boolean query is better at rank 100, but worse at rank 1000. This is because the second #uw50 (occurring in a window of size 50) operator in bold face was intended to be a #band operator. Using #uw50 must have filtered out many relevant documents at lower ranks after 100. Changing it back leads to better than keyword performance in infAP. Because the Boolean query paid no attention to the "porous particles" aspect, NDCG is slightly worse than the keyword query.

Cause: Human mistakes in formulating manual queries can easily occur, and are not easy to detect at lower ranks.

3.3 A Topic from ChemAthlon 2011

Topic title: *Manufacture of the potassium salt acesulfame-k*.

This topic came from patent search expert Stephen Adams, and was the worst performing topic within a total of four topics of the ChemAthlon task of PatOlympics 2011 [2]. This topic has very few relevant patents and very many false positives in the collection, because acesulfame-k is a very commonly used sweetener, which has been used to manufacture many different products, while the topic looks for the manufacture of the chemical itself.

Manual Boolean CNF style query used in ChemAthlon 2011:

```
#uw20(
    #syn(#1(acesulfame k) #1(3 oxathiazin 4 one))
    #syn(produce manufacture product)
)
```

Analysis: The original keyword query performed poorly, because first, there are different names that this popular sweetener has, second, the word "manufacture" mismatches most of the known relevant patents, and matches lots of the false positives that manufacture some product using the sweetener as an ingredient, and third, there are far more false positives than the few true relevant patents. Expanding the word "synthesis" to "manufacture" improved the retrieval performance significantly. However, the discovery of "synthesis" happened after the assigned 20 minutes of interaction time for the topic. During the competition, pressed by time, we only

reacted to the many false positives which were about manufacturing something else using the sweetener, and restricted the two concepts to appear in a 20-word window, which actually hurt performance by filtering out the few relevant patents returned by the original query.

3.4 Summary

False positives can be driven out of the rank list by enforcing more strict matching criteria (e.g. phrases or windowed occurrence), or can be driven to lower ranks by using CNF expansion to match and boost more relevant documents to the top. In practice, from the above three topics, it seems that the CNF expansion method is a better and more robust choice than restricting the results. In all the other four topics from the six Technology Survey topics, the manual queries did not use further windowed restriction. Three out of the four performed better than keyword, and in one case, the performance was very close to that of the keyword query.

4 Acknowledgements

We thank the organizer for all the help and efficient communication.

This work is supported by National Science Foundation grant IIS-1018317. The views and conclusions are the authors', and do not necessarily reflect those of the sponsor.

5 References

[1] INDRI - Language modeling meets inference networks. http://www.lemurproject.org/indri/. Retrieved Oct 24, 2011

[2] Mihai Lupu. PatOlympics - An Infrastructure for Interactive Evaluation of Patent Retrieval Tools. In *Proceedings of the Data infrastructurEs for Supporting Information Retrieval Evaluation - DESIRE 2011 Workshop*. 2011.

[3] Le Zhao. Modeling and Predicting Term Mismatch for Full-text Retrieval. *PhD Thesis Proposal*. Carnegie Mellon University. 2011.