BiTeM site report for the Claims to Passage task in CLEF-IP 2012

Julien Gobeill[1] and Patrick Ruch[1]

[1] BiTeM group, University of Applied Sciences, Information Studies, Geneva
{julien.gobeill,patrick.ruch@hesge.ch}

Abstract. In CLEF-IP 2012, we participated in the Claims to Passage task where the goal was to return relevant passages according to sets of claims, for patentability or novelty search purposes. The collection contained 2.3M of documents, corresponding to an estimated volume of 250M of passages. To cope with the problems induced by this large dataset, we designed a two-step retrieval system. In the first step, the 2.3M of patent application documents were indexed ; for each topic, we then retrieved the k most similar documents with a classical Prior Art Search. Document representations and tuning of the IR engine were set relying on training data and on the expertise we acquired in past similar tasks. In particular, we used not only claims for topics, but also the full description of the application document, and the applicants/inventors details ; moreover, we discarded retrieved documents that didn't share at least one IPC code with the topic. The k parameter ranged from 5 to 1000 according to the computed run. In the second step, for each topic (i.e. "on the fly"), we indexed the passages contained in these k most similar documents and queried with the topic claims in order to obtained the final runs. Thus, we dealt with approximately 11M of passages instead of 250M. The best k parameter with the training data was 10. Hence, we decided to submit four runs with k set to 10, 20, 50, and 100. Finally, we analyzed the training data and observed that the position of a passage in the document played a role, as passages at the end of the description were more likely to be relevant. Thus, we re-ranked each run according to passages' positions in the document in order to submit four supplementary runs.

Keywords. Information Retrieval; Intellectual Property

1 Introduction

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. We already took part in several evaluation campaigns on Information Retrieval (IR) in the Intellectual Property domain, such as previous CLEF [1], TREC[2], or NTCIR[3]. In CLEF-IP 2012, we participated in the Claims to Passage task where the goal was to return relevant passages from patents contained in the collection according to sets of claims, for patentability or novelty search purposes.

This task is known in computer science as Passage Retrieval, a subtask of Information Retrieval. We early identified two different strategies in order to retrieve the relevant passages: either a one-step retrieval, or a two-steps retrieval. The one-step retrieval consists in building a unique search engine by indexing all passages. The two-steps retrieval consists in building a first unique search engine by indexing all documents, then, for each topic (i.e. „on the fly"), building a second search engine by only indexing passages belonging to the retrieved documents. The CLEF-IP 12 collection contained 2.3M of documents, corresponding to an estimated volume of 250M of passages. To cope with the problems induced by this large dataset, we chose the two-steps retrieval strategy, as described in Fig.1.

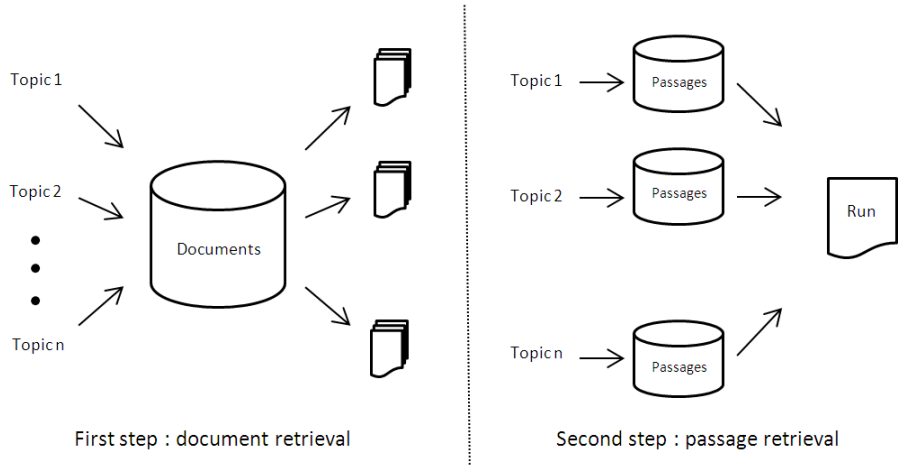

Fig. 1. The two-steps strategy we investigated.

2 Strategies

In the first step, the 2.3M of patent application documents were indexed ; for each topic, we then retrieved the most similar documents with a classical Prior Art Search. Document representation and tuning of the IR engine were set relying on training data and on the expertise we acquired in past similar tasks. For document representation, we used titles, abstracts, claims, applicants and inventors details, and IPC codes (complete format, e.g. G09G 3/28). For topic representation, we exploited the provided patents and also used titles, abstracts, claims, applicants and inventors details, and IPC codes, along with the full description sections. We thus obtained, for each query, a set of retrieved patents. Then, we applied a supplementary post-processing strategy investigated in previous CLEF-IP, by discarding retrieved patents that did not share at least one IPC code with the topic patent.

In the second step, for each topic, we extracted passages contained in the *k* first retrieved patents. Then, for each topic, we indexed these passages and queried only with

the claims provided in the topic in order to obtain our runs. Relying on training data, we evaluated different values of k, ranging from 5 to 1000.

Indexing and Retrieval were computed with the Terrier platform [4], which is designed for large collections such as TREC or CLEF collections. We chose settings which proved to be efficient in the past competitions: PL2 as weighting scheme, and Bo1 as Query Expansion model, both with default parameters and with Porter stemming [5]. For multilingual purposes, we simply chose to only index English sections, and to use Google in order to translate the topics from French or German into English.

Finally, we analyzed the qrel provided with the training data, focusing on the position (within the description section) of the relevant passages. We thus divided, for each patent, the description section into ten equal parts. Then, we analyzed from which part came the passages contained in the qrel, and the passages contained in our run (for k=10). $P_{qrel}(i)$ is the percentage of relevant passages in the qrel that belong to the i-th part of the description, i ranging from 1 (the beginning) to 10 (the end). $P_{run}(i)$ is the percentage of relevant passages in our run that belong to the i-th part of the description Fig.2 illustrates these distributions.

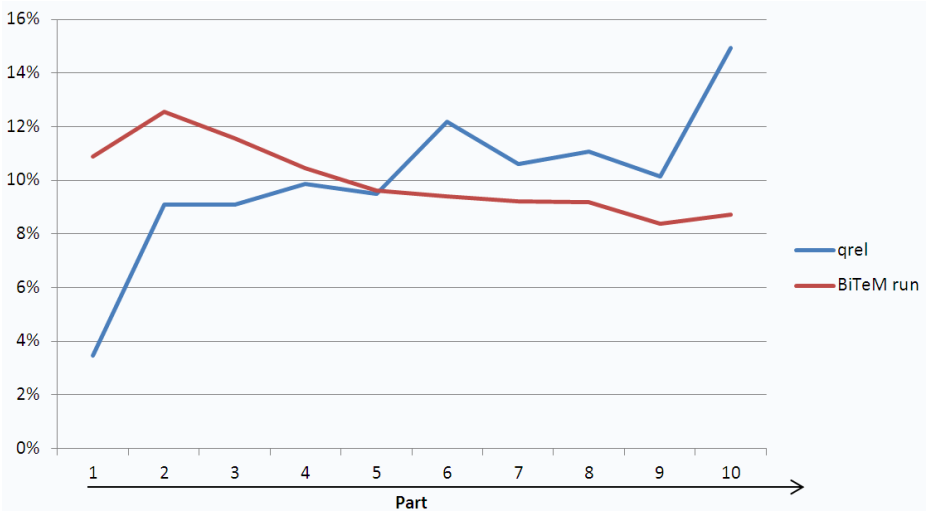

Fig. 2. Distribution of passages, according to their position in the description section, in the training data qrel and one of our runs (for k=10).

Both distributions are opposite. In the qrel, the passages belonging to the end of the description are more likely to be relevant. On the contrary, our search system tends to favor passages belonging to the beginning of the description. Hence, we computed weights $W(i)$ for each part according to the qrel distribution, by dividing $P_{qrel}(i)$ by $P_{run}(i)$, then we re-ranked our runs by boosting scores according to these weights. This re-ranking strategy is obviously applied only to passages belonging to the description section.

For evaluations, we computed Mean Reciprocal Rank (MRR), which is the multiplicative inverse of the rank of the first correct returned answer [5].

3 Results on training data

First, we evaluated different values of *k* with the training data. Results are presented in Tab. 1. It appears that the best value for *k* is 10. *k*=10 means that, for each query, we retrieved passages only within the 10 first retrieved patents.

k	MRR
5	0.014
10	0.017
20	0.01
50	0.013
100	0.013
200	0.007
500	0.004
1000	0.003

Table 1. Results obtained with the training data, in terms of MRR, according to the *k* value

Finally, we evaluated the impact of our re-ranking strategy with training data, and observed a slight improvement in terms of MRR, ranging from +2% to +6% according to the value of *k*.

4 Conclusion

Hence, we decided to submit four official runs with different values of *k*: 10, 20, 50 and 100. As participants were allowed to submit up to 8 runs, we applied our re-ranking strategy to all the mentioned runs in order to obtain four supplementary official runs.

5 References

1. Gobeill,J., Pasche,E., Teodoro,D., Ruch,P.: Simple pre and post processing strategies for partent searching in CLEF Intellectual Property Track 2009. Proceedings of CLEF (2009)
2. Gobeill,J., Gaudinat,A., Pasche,E., Teodoro,D., Vishnyakova,D., Ruch,P.: BiTeM site report for TREC Chemistry 2010: Impact of Citations Feedback for Patent Prior Art Search and Chemical Compounds Expansion for Ad Hoc Retrieval. Proceedings of TREC (2010)
3. Teodoro,D., Gobeill,J., Pasche,E., Ruch,P.: Report on the NTCIR 2010 Experiments: automatic IPC encoding and novelty detection for effective patent mining. Proceedings of NTCIR (2010)
4. Ounis,I., Lioma,C., Macdonald,C., Plachouras,V.: 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)
5. Porter,M.: An algorithm for suffix stripping. Program, vol. 14, pp. 130-137 (1980)
6. Voorhees,E.: Overview of the Question Answering Track. Proceedings TREC (2001).