Patent Prior-art search Failure Analysis

Why Patent Prior-art Search Fails?

Sean Fogarty
NICTA & ANU
Canberra, Australia
name.surname@nicta.com.au

G.K.M. Tobin
NICTA & ANU
Canberra, Australia
name.surname@nicta.com.au

Lars Thørväld
NICTA & ANU
Canberra, Australia
name.surname@nicta.com.au

ABSTRACT

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query Formulation

General Terms

Theory

Keywords

patent search, Query Reformulation, Data Analysis

1. INTRODUCTION

2. BASELINE

3. TERM ANALYSIS

The main complain about patent search is insufficient match between the content of patent queries and relevant patents[1, 2]. However, our analyses showed that only %20 overlap is sufficient for the system to retrieve a relevant or non-relevant patent at top-100 and except for few queries, non-retrieved relevant patents had enough matched term with the query. So, we start our experiments with term analysis for patent query and retrieved documents.

3.1 Discriminative Words

For our initial experiments, we identified the *discriminative words* by positive scoring the words in relevant documents and negative scoring the irrelevant one.

$$score(t, Q) = Rel(t) - Irr(t) \quad (1)$$

$$t \in \{\text{terms in top-100 retrieved documents}\}$$

Where $Rel(t)$ is the average term frequency in retrieved relevant patents and $Irr(t)$ is the average term frequency in retrieved irrelevant patents. Words with a positive score consider *useful words* since they are more frequent in relevant patents while the words with negative score are *noisy words* as they appeared more frequently in irrelevant patents.

Surprisingly, we could not find any correlation between the percentage of *useful words* and the performance. We expected a higher performance for the queries with more *useful words*.

3.1.1 Optimal RF[1] Query Formulation

We hypothesize that a query formulated by *useful terms* is optimal since they are all frequent in relevant patents and rare in irrelevant ones. Table 1 compares the performance for baseline where the query is the full patent query both weighted and unweighed with the performance for optimal RF query weighted with the score of each term(formula 1) and unweighed.

Table 1: .

	Pat.Query Weight:TF	Pat.Query Weight:1	Opt.RFQuery Weight:Score(t)	Opt.RFQuery Weight:1
PRES	0.5355	0.4268	0.6086	0.6087
MAP	0.1618	0.1181	0.4617	0.5075
A. Recall	0.5491	0.4385	0.6129	0.6118

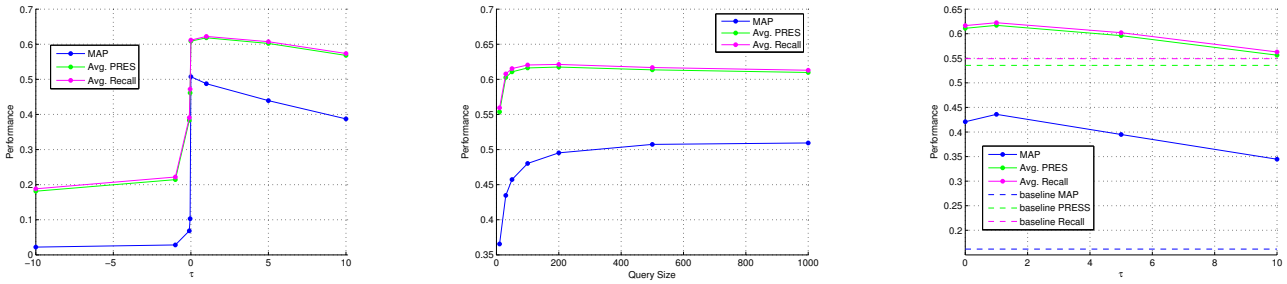
It can be seen that 'MAP' jumps from 0.1618 to **0.5075** which is about %35 increase. We use a score threshold(τ) to formulate the RF query(we select the terms with $score(t) > \tau$). Fig. (1-a) indicates two important facts. First, it shows that increasing the threshold results in the lower performance. Second, the system is over-sensitive to the *noisy words*($\tau < 0$). Fig. (1-b) shows that formulating a query with up to 200 *useful words* helps performance whereas the performance improves slightly by adding more than 200 words.

3.1.2 Query Reduction by Relevance Feedback

Our experiments led us to another hypothesis that a patent query contains sufficient words matched with the relevant patents and the *noisy words* are the main cause of the low effectiveness. Therefore, we use RF *useful terms* to reduce the patent query terms by selecting terms such that: $t \in \{Q \cap (useful\ terms)\}$. Fig. (1-c) explicitly shows that a patent query contains sufficient words to retrieve relevant patents at top of the list. We only need to keep the *useful terms* and prune out the *noisy words*.

[1]Relevance Feedback

SIGIR '15, August 9-13, 2015, Santiago, Chile



(a) (b) (c)

Figure 1: How score threshold(τ) and query size controls the performance. (a) Performance versus the score threshold. (b) Performance versus the query size. (c) System performance when we reduced the query by RF: $query = Q \cap (useful\ terms)$, where Q is the patent query and $useful\ terms = \{t|score_{RF}(t) > \tau\}$.

3.2 What did not Work

We achieved a high 'MAP' by relevance feedback, however, it will cost to have relevance feedback since our users are professional patent examiners and recognizing a relevant patent is a time-taking and demanding task. So, we tried ways to refine the best query out of the patent query.

3.2.1 *Identify the Noisy Words*

First, we attempted to identify the noisy words. We hypothesized that the noisy words are frequent in top-100 retrieved documents. So, we calculate the average term frequency of each term as follows:

$$score_{DF}(t) = \frac{1}{100} \sum_{t \in \{Top-100\}} TF(t) \qquad (2)$$

where TF(t) is term frequency of each term in top-100 retrieved patent document.

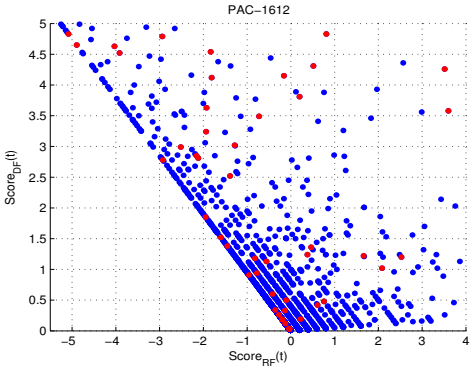


Figure 2: Anecdotal example: Scatter plot of RF score and Document Frequency(DF) score of words in top-100 retrieved documents. Each blue point is a vocabulary in top-100 retrieved document vocabulary set. Red point are query words with term frequency higher than 5($QTF(t) > 5$).

3.2.2 *Pseudo Relevance Feedback*

3.3 Improvement by Minimum User Effort

We could not improve the effectiveness without accessing the relevance feedback. In this experiment, we select query terms using only the first relevant patent document based on the hypothesis that a patent examiner can recognize the first relevant patent with minimum effort at top-5. Table 2 shows that we can double the 'MAP' by using only the first-ranked relevant document.

Table 2: System performance when only the first relevant patent used for query reduction. τ is RF score threshold, and k indicates the number of first relevant retrieved documents.

	$(k = 1$ & $\tau = 1)$	$(k = 3$ & $\tau = 0)$	$(k = 3$ & $\tau = 1)$
PRES	0.5016	0.5699	0.5727
MAP	0.3040	0.3879	0.3872
A. Recall	0.5090	0.5757	0.5787

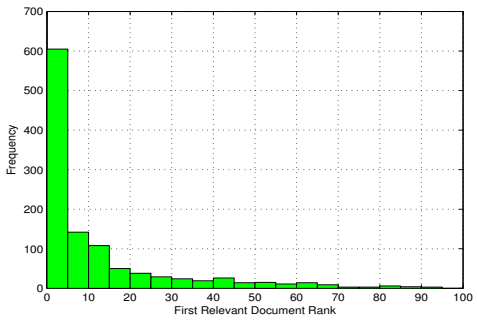


Figure 3: The distribution of the first relevant document rank over test queries which have TPs

4. RELATED WORK

5. CONCLUSIONS

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] W. Magdy. *Toward higher effectiveness for recall-oriented information retrieval: A patent retrieval case study*. PhD thesis, Dublin City University, 2012.

[2] P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 113–122. ACM, 2013.