Facts behind the Words of a Patent Query

Why Patent Prior-art Search Fails?

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

This result explicitly demonstrates that the *noisy words* are the main cause of the low effectiveness not the insufficient term match.

3.2 What did not Work

3.2.1 Pseudo Relevance Feedback

3.2.2 Identify the Noisy Words

[1]Relevance Feedback

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

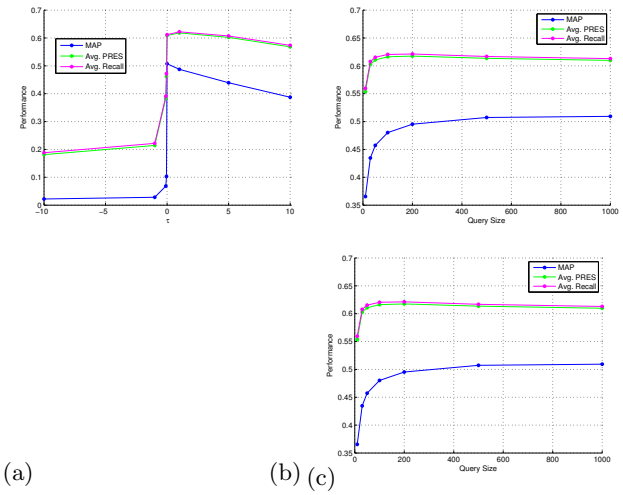

Figure 1: How score threshold(τ) and query size controls the performance. (a) Performance versus the score threshold. (b) Performance versus the query size.

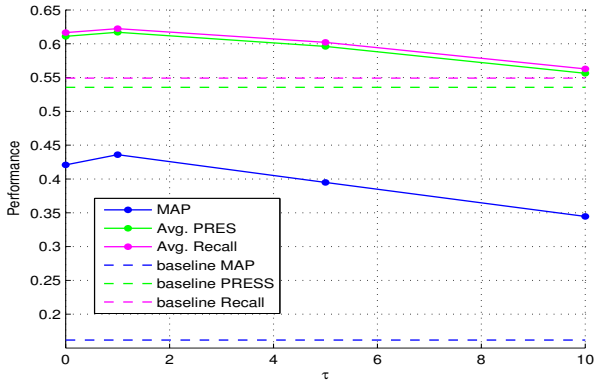

Figure 2: Optimal patent query performance vs. score threshold of useful terms to select patent query terms. $query = Q \cap (useful\ terms)$, where Q is the patent query and $useful\ terms = \{t|score_{RF}(t) > \tau\}$..

3.3 Improvement by Minimum User Effort

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.