Necessary and Frequent Terms in Queries

Jiepu Jiang
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts Amherst
jpjiang@cs.umass.edu

James Allan
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts Amherst
allan@cs.umass.edu

ABSTRACT

Vocabulary mismatch has long been recognized as one of the major issues affecting search effectiveness. Ineffective queries usually fail to incorporate important terms and/or incorrectly include inappropriate keywords. However, in this paper we show another cause of reduced search performance: sometimes users issue reasonable query terms, but systems cannot identify the correct properties of those terms and take advantages of the properties. Specifically, we study two distinct types of terms that exist in all search queries: (1) necessary terms, for which term occurrence alone is indicative of document relevance; and (2) frequent terms, for which the relative term frequency is indicative of document relevance within the set of documents where the term appears. We evaluate these two properties of query terms in a dataset. Results show that only 1/3 of the terms are both necessary and frequent, while another 1/3 only hold one of the properties and the final third do not hold any of the properties. However, existing retrieval models do not clearly distinguish terms with the two properties and consider them differently. We further show the great potential of improving retrieval models by treating terms with distinct properties differently.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, retrieval models.*

General Terms

Performance, Experimentation, Human Factors.

Keywords

Query; term frequency; term occurrence.

1. INTRODUCTION

Term frequency (TF) is widely used as an important heuristic in retrieval models [1–3]. The assumption is that documents with comparatively higher frequencies of query terms are more likely to be relevant. However, we suspect that in many cases this assumption does not hold. Instead, users may adopt some query terms to simply include or exclude documents regardless of the occurrences of the terms – that is, in those cases TF does not indicate the relevance of documents as long as the terms appear. In such cases, retrieval models that heavily exploit TF may incorrectly rank some non-relevant documents with high frequencies of the terms to the top.

We define the following two properties of query terms. We say that a term is *necessary* to a topic if most relevant documents contain the term. Documents with no occurrences of the necessary term are unlikely to be relevant. In comparison, we say that a term is *frequent* to a topic if relevant documents usually have relatively more occurrences of the term comparing to the non-relevant ones. Documents in which the frequent term appears many times are more likely relevant compared to those where the term appears less frequently. Note that the two properties do not conflict with each other: a term can be both necessary and frequent.

We hypothesize that both necessary and frequent terms exist in user queries, but some query terms may only conform to one of the two properties. We study the following research questions:

RQ1: Do query terms differ with respect to the two properties? We examine the two properties of query terms in a dataset based on term occurrences in relevant and non-relevant documents.

RQ2: How do users perceive the two properties of query terms? Do users' opinions agree with those learned from the dataset and do users agree with each other? We ask assessors to annotate query terms regarding the two properties and analyze the results.

RQ3: Assuming we know the properties of query terms, can we improve search performance by treating terms differently? We show a simple approach that can achieve 35% improvement in nDCG@10 compared to the query likelihood model, if it knows these properties of query terms. Results suggests great potential for improving search performance by identifying properties of query terms and treating them differently in retrieval models.

2. EVALUATION OF TERM PROPERTIES

In this section, we define indicators and examine query term properties in the TREC Robust 2004 dataset.

2.1 Indicators of Term Properties

We denote the degree to which a query term w is necessary to a topic by $P(X=1|R)$, the probability of observing w in the set of relevant documents, R. X=1 refers to the occurrence of w in a document regardless of its frequency. In a dataset with R being judged, we can estimate $P(X=1|R)$ by Equation (1), where: N is the total number of documents in R; N_w is the number of documents in R where w appears at least once. The greater the value of $P(X=1|R)$, the more necessary the term to the topic.

$$\hat{P}\left(X=1|R\right)=\frac{N_w}{N} \quad (1)$$

We evaluate to what degree a query term w is frequent to a topic by comparing $P(w|R)$ and $P(w|NR)$, where NR is the set of non-relevant documents. $P(w|R)$ is the probability of the term w in relevant documents, which is estimated by Equation (2), where: $P(w|d)$ is the probability of w in the multinomial document language model of d; each document d in R has an equal weight 1/N to contribute to $P(w|R)$. We estimate $P(w|d)$ using maximum likelihood estimation with Dirichlet smoothing [4]. The parameter μ is selected to optimize the nDCG@10 of query likelihood model

SIGIR'14, July 06–11, 2014, Gold Coast, QLD, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2257-7/14/07 $15.00.
http://dx.doi.org/10.1145/2600428.2609536

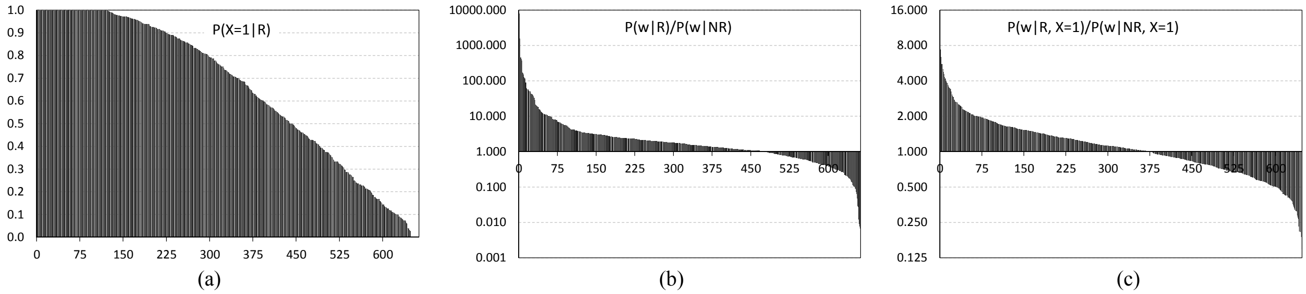


Figure 1. Distribution of *P*(*X*=1|*R*), *P*(*X*=1|*R*)/*P*(*X*=1|*NR*), *P*(*w*|*R*)/*P*(*w*|*NR*), and *P*(*w*|*R*, *X*=1)/*P*(*w*|*NR*, *X*=1) on 663 query terms.

in the dataset. We estimate $P(w|NR)$ in a similar form but among the set of non-relevant documents. The greater the value of $P(w|R)$ compared to $P(w|NR)$, the more *frequent* is the term *w*.

$$\hat{P}(w \mid R)=\frac{1}{N}\cdot\sum_{d\in R}P(w \mid d) \qquad (2)$$

It should be noted that we can easily observe $P(w|R) > P(w|NR)$ when *w* is necessary for *R* but rarely appears in *NR*. Therefore, we further examine a stronger form of the frequent term property: within the set of documents where *w* appears at least once, relatively higher frequency of the term indicates greater likelihood of relevance. We quantify this stronger property by comparing $P(w|X=1,R)$ and $P(w|X=1,NR)$. The two probabilities are estimated similar to Equation (2), but within the set of relevant and non-relevant documents where *w* appears at least once.

2.2 Evaluation

We calculate the indicators related to term properties in TREC Robust 2004 dataset. The dataset includes 250 queries and 663 query terms (counting multiple occurrences of the same term in different queries). We remove the Indri standard stopwords and stem using the Krovetz stemmer when processing documents and queries. Figure 1(a), 1(b), and 1(c) show the distribution of $P(X=1|R)$, $P(w|R)/P(w|NR)$, and $P(w|R,X=1)/P(w|NR,X=1)$ for the 663 query terms.

Results show that it is very common to use query terms that do not hold the two properties. As shown in Fig. 1(a), among the 663 query terms, only 18.5% are fully necessary – i.e., $P(X=1|R)=1$ – and 44.8% roughly hold the necessary property – $P(X=1|R)\geq 0.8$. Moreover, 33% of the query terms do not hold the necessary property ($P(X=1|R)<0.5$), and 50% of the queries have at least one such term. Using query terms with the frequent term property is also very common in the dataset: Figures 1(b) and 1(c) show that 475 out of the 663 query terms (71.6%) hold the basic frequent term property, but only 373 (56.3%) hold the stronger form where $P(w|R,X=1)/P(w|NR,X=1)>1$. Among the 250 queries, 57.8% have at least one term that does not hold the frequent term property and 75.1% have at least one term that does not hold the stronger form of the frequent term property.

We further evaluate the relation between search effectiveness and using query terms that do not hold the two properties. Figure 2 shows the average nDCG@10 of queries in which at least one term's value of the three indicators is less than *P*, where *P* ranges from 0.1 to 1.0. Results suggest that queries with terms that do not hold either of the two properties are less effective. For example, for the set of queries with at least one term's value of $P(X=1|R) <$ 0.5, the nDCG@10 of these queries is only 0.356, less effective than those of the 250 queries on average. For queries with terms that do not hold either of the two properties, search performance declined by a greater magnitude. However, we noticed that for queries with terms that have $P(w|R)/P(w|NR) < P$ ranging from 0.2 to 0.6, there are no apparent differences in the queries' search performance. This indicates that $P(w|R)/P(w|NR)$ is less indicative of term's search effectiveness. In following discussions, we use the stronger form of frequent term property and adopt $P(w|R,X=1)/P(w|NR,X=1)$ as the indicator.

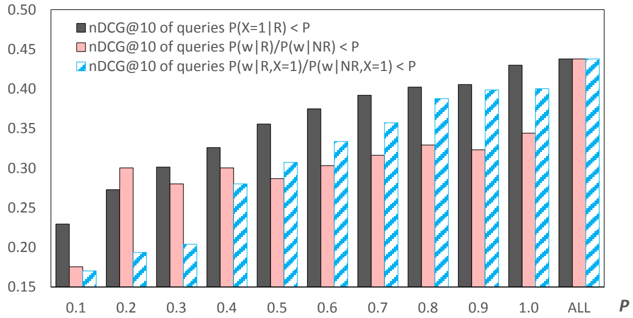
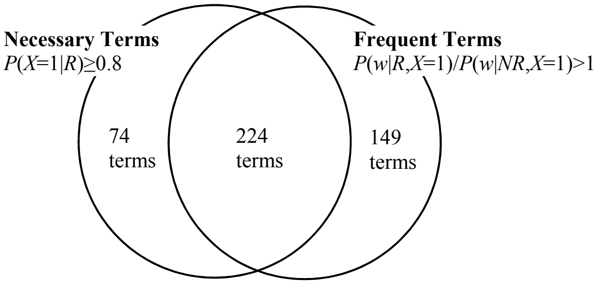


Figure 2. nDCG@10 of queries with at least one term for which the three indicators < *P*. *P* ranges from 0.1 to 1.0. "ALL" shows the average nDCG@10 of ALL queries.



216 terms are neither necessary nor frequent.

Figure 3. Overlap of terms conforming to two properties.

We further show the overlap of query terms conforming to the two properties in Figure 3. Among the 663 terms, 224 (33.8%) are both necessary and frequent, but 223 (33.6%) only hold one of the two properties. The remaining 216 terms (32.6%) are neither necessary nor frequent. This suggests different strategies should be adopted to improve ineffective queries. The 216 terms that do not hold either property are not indicative of document relevance and would be better removed. For the 74 terms only having the necessary term property, we should prefer documents where the term appears but do not give further credit to high term frequency. For the 149 terms only having the frequent term property, we should prefer documents where the term appears many times over those where the term appears only once or twice, but it may be risky to filter out documents without any occurrence of the term.

To summarize, our results show that whether or not a term holds the two properties affects the search effectiveness of queries. In the dataset, only 1/3 of the query terms hold both properties. Another 1/3 hold only one of the two properties. The other 1/3 have neither property. This suggests that we may improve search systems in two different ways: identify terms with these properties and adopt different ranking criteria; predict terms without any of the two properties and discount the effects of such terms in ranking. In further sections, we explore such potentially by assuming we can correctly identify properties of terms.

3. USER JUDGMENTS OF PROPERTIES

In this section, we study whether or not users can make correct judgments of these two term properties. This is meaningful for two reasons. First, if the users make poor judgments on query properties, it provides a new explanation for ineffective queries. Second, it the users can make correct judgments, systems may benefit from providing query languages allowing users to express their sense of the properties.

We asked 10 users to annotate 100 TREC queries selected from the TREC Robust 2004 dataset (Topic 301-400). Each user annotated 15 queries, with 10 overlapping with another two users. For example, the first user annotated query 301-315, the second user on query 311-325, … , and the last user on query 391-400 as well as 301-305. This resulted in 10 users' annotations on the 100 queries. For 50 queries, we have only one user's annotation, and for the other 50, we have two users' annotations, so that we can study users' agreements on the properties of query terms. For each query term, we asked users two yes/no questions as follows. We say that a user annotated a query term as necessary or frequent if the answer on Q1 or Q2 is yes, respectively.

Q1: I believe most of the relevant results should have this word. Results that do not contain this word are unlikely to be useful.

Q2: I believe this word should appear many times in relevant results. Results in which the word appears only once or twice are less likely to be useful.

We found that pairs of users have some agreement on whether or not a term is necessary, but their opinions are rather independent of each other on the frequent terms. Among the 126 query terms involving two users' annotations, users agreed in 67% of the cases regarding whether or not a term is necessary. However, they agreed only in 48% of the cases on whether a term is frequent.

Table 1. Correctness of user annotation of term properties.

Property	P	Num Y/N by P	Num Y/N by Users	User Acc / Prior	Class	Prec	Rec
Necessary	0.8	88/164	201/51	0.50/**0.65**	Y	0.41	0.93
					N	0.88	0.27
	0.5	145/107	201/51	**0.63**/0.57	Y	0.63	0.87
					N	0.63	0.30
Frequent	1.0	124/128	165/87	**0.60**/0.51	Y	0.57	0.76
					N	0.66	0.45
	0.8	156/96	165/87	0.61/**0.62**	Y	0.67	0.71
					N	0.48	0.44

Table 1 shows the accuracy of users' annotations of query term properties comparing to those evaluated by the values of $P(X=1|R)$ and $P(w|R,X=1)/P(w|NR,X=1)$. Results show that in general it is difficult for users to make correct judgments on the query terms' properties. If we use $P(X=1|R)>0.5$ as the criteria for necessary terms, users' judgments are slightly better than a classifier using prior probability of the classes (accuracy 0.63 versus 0.57). When we use $P(w|R,X=1)/P(w|NR,X=1)>1.0$ as the threshold for frequent terms, users did also only slightly better than a classifier using prior probabilities (accuracy 0.60 versus 0.51). The accuracy and precision of user judgments look not useful. Moreover, when we adopt different criteria for term properties, e.g. $P(X=1|R)>0.8$, users' judgments may even be worse than a classifier using the prior probability of classes.

To conclude, the results of user annotation on query term properties show that it is very difficult for users to select the properties of query terms prior to looking at search results. Users also agree only slightly with others on whether a term property applies. Specifically, users' judgments on frequent terms are completely independent of others.

4. SYSTEMS USING TERM PROPERTIES

In this section, we explore the potential of improving retrieval systems assuming we know the properties of terms correctly. The prediction of term properties is left for future work.

4.1 Approaches

Let q be a query. We assume we know the set of necessary terms q_N and the set of frequent terms q_F. Note that q_N and q_F can be empty set, and a term in q may be in neither q_N nor q_F. We rank a document d by Equation (3), where: we assume q_N and q_F are independent given d; each term in q_N and q_F are generated independently of other terms from d by different process $P_N(w|d)$ and $P_F(w|d)$.

$$\begin{aligned} &P(q_N, q_F \mid d) \\ &= P(q_N \mid d) P(q_F \mid d) \\ &= \prod_{w \in q_N} P_N(w \mid d) \cdot \prod_{w \in q_F} P_F(w \mid d) \end{aligned} \tag{3}$$

We calculate $P_N(w|d)$ and $P_F(w|d)$ in Eq(4) and Eq(5). In Eq(4), we calculate $P_N(w|d)$ as the probability of selecting a term w from d's vocabulary V_d ignoring the frequency of terms in d. $|V_d|$ is the size of d's vocabulary. $P_N(w)$ is the probability that w exists in a the vocabulary of a document in the whole corpus. In a corpus of k documents, we estimate $P_N(w)$ as Eq(6). μ_N is a parameter for smoothing. $P_F(w|d)$ is simply the probability of a term w from the multinomial document language model of d, estimated using maximum likelihood estimation with Dirichlet smoothing. In our experiments, we set μ_F to the value that can maximize nDCG@10 of using all terms as q_F and no term as q_N for retrieval (equivalent to query likelihood model). In contrast, we set μ_N to the value that can maximize nDCG@10 of using all terms as q_N and no term as q_F for retrieval.

$$\hat{P}_N(w \mid d) = \frac{1 + \mu_N \cdot P_N(w)}{|V_d| + \mu_N} \tag{4}$$

$$\hat{P}_F(w \mid d) = \frac{c(w,d) + \mu_F \cdot P_F(w)}{|d| + \mu_F} \tag{5}$$

$$\hat{P}_N(w) = \frac{1}{k} \cdot \sum_d \frac{1}{|V_d|} \tag{6}$$

For a necessary term w in q_N, $P_N(w|d)$ totally ignores the frequency of w in d. Its value depends only on whether or not w appears in d. In addition, it favors documents with a small vocabulary. (This is intuitively correct because observing w in d is less informative if d is very long and has a large vocabulary.) When we put all the query terms into q_F and none into q_N, Equation (3) falls back to the query likelihood language model.

4.2 Search Effectiveness

In this section, we evaluate the approaches proposed above by assuming different sets of necessary and frequent terms. Table 2 shows the results. For "q_N" and "q_F" in Table 2, "none" means do not use any terms, "all" means using all query terms, and "best" means using the best possible combination of query terms (the set of query terms that leads to the best nDCG@10).

We first evaluate the effectiveness of $P_N(w|d)$ and $P_F(w|d)$ on different set of terms individually. Unsurprisingly, using all terms as necessary terms (N++) performs worse than using all terms as frequent terms (F++ and also Query Likelihood). However, simply ignoring term frequencies of all documents still achieved nDCG@10 as high as 0.293. This indicates that solely considering term occurrences is still useful in many cases. However, simply using all terms as both necessary and frequent terms (N++F++) did not result in any improvements.

We further examine whether removing inappropriate terms from q_N or q_F can lead to improved search performance. As shown in Table 2, removing inappropriate terms from q_F can potentially improve nDCG@10 from 0.438 (F++) to 0.514 (F+), and from 0.436 (N++F++) to 0.528 (F++F+). Similarly, removing terms from q_N can potentially improve nDCG@10 from 0.293 (N++) to 0.329 (N+), and from 0.436 (N++F++) to 0.503 (N+F++). When we remove inappropriate words from both q_N and q_F (N+F+), we can potentially improve nDCG@10 to 0.590, which is about 35% improvements comparing to QL and N++F++. This suggests that there is great potentiality of improving search performance if we can predict correctly the frequent and necessary words.

However, it should be noted that the best set of terms for q_N and q_F are dependent of each other. When we use the best set of q_N in N+F++ and the best set of q_F in N++F+ for retrieval (N+F+ local), there will be 10% decline of nDCG@10 comparing to N+F+. Besides, we found that a part of the improvement of search performance comes from removing inappropriate terms from both q_N and q_F. If we restrict that all the query terms should be in at least one of q_N and q_F (N+F+ (-rmv)), the nDCG@10 declined from 0.590 to 0.552, although still a substantial improvement comparing to F++ (QL).

We further examine whether using the indicators of properties in section 2, i.e., $P(X=1|R)$ and $P(w|R,X=1)/P(w|NR,X=1)$, can effectively select the appropriate set of terms for q_N and q_F to enhance search performance. We examined a simple rule-based approach as follows. We start with all query terms in q_F and no terms in q_N. We remove terms in q_F if $P(w|R,X=1)/P(w|NR,X=1) < 1.05$. If the removed term has $P(X=1|R)>0.2$, we add the term into q_N. Besides, we add all terms with $P(X=1|R)>0.95$ into q_N. This simply rule-based approach (N+F+ P) improves nDCG@10 by 8.7% comparing to F++ (using all terms for q_F). This suggests that the two indicators are effective criterion of selecting q_N and q_F. However, the performance of the selected q_N and q_F cannot be compared with the best possible q_N and q_F in N+F+. This indicates that the two indicators are not enough for selecting q_F and q_N. The exploration of predictors for q_F and q_N is left for future works.

Earlier, we showed that users made poor judgments on the properties of query terms. To further verify the quality of users' judgments, we select terms into q_N and q_F if users answered yes in Q1 and Q2. As shown in Table 2, this approach reduces search performance. The nDCG@10 is 0.416 (N+F+ user) versus 0.443 in QL on the same set of 100 queries. This further confirms that it is difficult for users to make useful judgments on term properties.

Table 2. Potential improvements of search performance.

Label	q_N	q_F	**nDCG@10**	**Change / Baseline**
F++ (QL)	none	all	0.438	-
F+	none	best	0.514	+17.4% / F++
N++	all	none	0.293	-
N+	best	none	0.329	+12.3% / N++
N++F++	all	all	0.436	-
N++F+	all	best	0.528	+21.1% / N++F++
N+F++	best	all	0.503	+15.4% / N++F++
N+F+	best	best	**0.590**	+35.3% / N++F++
N+F+ local	best.L	best.L	0.541	+24.1% / N++F++
N+F+ (-rmv)	best	best	0.552	+26.6% / N++F++
N+F+ P	$P(X\mid R)$	$P(w\mid R,X)$ / $P(w\mid NR,X)$	0.476	+8.7% / QL
N+F+ user	user	user	0.416	QL: nDCG@10 0.443 (100 queries)
F+RM	none	RM100	0.644	-

* N/F in the run labels refers to q_N/q_F; ++ means using all terms; + means using selected query terms.

So far we limit the set of query terms among those being issued by the users, and the improvements of search performance mainly comes from correct identification of the necessary terms and the frequent terms. We compare our approach with query expansion on the potential of improving search performance. We estimate the true relevance model based on qrels, and use the top 100 terms ("RM100") as q_F for search. As shown in Table 2, solely working on the set of query terms issued by users, N+F+ is not much worse than F+RM (true relevance model) on nDCG@10, which extensively exploits the representative terms in relevant results.

5. FUTURE WORK

In this preliminary study, we show that retrieval models that exploit term frequency can potentially be improved substantially by separately considering TF for some query terms and counting only occurrence or non-occurrence for some other query terms. This conclusion comes from our findings that query terms hold different properties. Specifically, sometimes the frequencies of terms do not indicate document relevance as long as the terms appear. In such cases, existing retrieval models may incorrectly rank documents with high term frequencies to the top. Queries with terms lacking either property are less effective in general.

Future work on this topic mainly focuses on the prediction of an appropriate set of terms in q_N and q_F. As discussed in section 4, though values of the two indicators can effectively predict q_N and q_F, it is far from perfect and the two indicators are also computed based on known relevance judgments.

Our study is closely related but different from the recent work of term necessity prediction by Zhao and Callan [5, 6]. Zhao et al. focused on predicting $P(w|R)$ and aimed at solving term mismatch by selecting terms with highly predicted $P(w|R)$ values for query expansion. In comparison, we do not expand the query but aim at recognizing the correct properties of query terms that are issued by the users. The two approaches follow different directions but may potentially be combined. As shown in Table 2, our approach may have substantial improvements on search performance that is comparable to those can be achieved by predicting $P(w|R)$.

ACKNOWLEDGEMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] Lafferty, J. and Zhai, C. 2001. Document language models, query models, and risk minimization for information retrieval. In Proc. SIGIR'01: 111-119.

[2] Ponte, J.M. and Croft, W.B. 1998. A language modeling approach to information retrieval. Proc. SIGIR'98: 275-281.

[3] Robertson, S.E. et al. 1995. Okapi at TREC-3. NIST Special Publication 500-226: Proceedings of the Third Text REtrieval Conference (TREC-3).

[4] Zhai, C. and Lafferty, J. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In Proc. SIGIR'01: 334–342.

[5] Zhao, L. and Callan, J. 2012. Automatic term mismatch diagnosis for selective query expansion. In Proc. SIGIR'12: 515-524.

[6] Zhao, L. and Callan, J. 2010. Term necessity prediction. In Proc. CIKM'10: 259–268.