Word importance discrimination using context information

Danil Nemirovsky[a,b] and Vladimir Dobrynin[b]

[a] *INRIA Sophia Antipolis, France*

[b] *St. Petersburg State University, Russia*

danil.nemirovsky@gmail.com, v.dobrynin@bk.ru

Abstract

Word importance discrimination is a task deserving attention when one treats a topic from TREC where a topic is quite long. The goal of the process is to estimate importance of words which carry any (additional) information about user information needs. In our experiments we estimated word importance using context information of a word.

Key words: Word importance discrimination, Context, Clustering, CDC

1 Introduction

Word importance discrimination is a task deserving attention when one treats a topic from TREC where a topic is quite long. The goal of the process is to estimate importance of words which carry any (additional) information about user information needs. Word importance discrimination task is strongly related to word filtering where word importance is a binary value. There were proposed several approaches addressing word filtering. Luhn [3] uses a simple filter based on frequency of term occurrence, Bookstain et al. [1] detect content-bearing words by serial clustering, Picard [4] suggests to use term similarities, and Takayama et al. [6] used SVD decomposition of co-occurrence matrix. In our experiments we estimated word importance based on context of a word, which is a probability distribution over words that can be met in the same documents as the given word. Intuitively, one can say that a word is important if it has specific meaning in a domain and occurs in the relatively small number of documents. This implies that a word has specific meaning if its context is of low entropy. We use this idea to estimate importance of both

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Word importance discrimination using context information	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **INRIA,Sophia Antipolis, France,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**7**	

words and phrases in a document collection and perform experiments to find out what kind of word importance discrimination shows better results.

2 Methodology

2.1 Context Document Clustering algorithm (CDC)

Context Document Clustering algorithm is a scalable clustering algorithm which full description of can be found in [2, 5].

In our TREC2008 experiment we use idea of term context which plays an important role in CDC algorithm. Let each document of a collection be presented as a probability distribution over the set of all the terms of the collection called a profile of the document. A document is presented by a probability distribution over the set of terms in the model:

$$p(t|d) = \frac{tf_{t,d}}{N_d}, \tag{1}$$

where $tf_{t,d}$ is the number of occurrence of term t in document d and N_d is the total number of terms in document d. The occurrence of a term in a document is assumed independent from all other terms of the document. Nothing is assumed about the notion of "term" except the fact that a document consists of terms and the set of all the terms of the collection is the set of terms met in a document of the collection. A context is created for each term which is not very common (e.g. it is an upper bound for the document frequency of the term) or very rare (e.g. it is a lower bound for the document frequency of the term) in the collection. A context of a term t is a probability distribution over all the terms in the collection and the entry of the distribution is probability to meet the term t with another term in the same document.

The term contexts are used to select important words (more precisely, important terms) from the dictionary of the collection. Intuitively, one can say that a word is important if it has specific meaning in a domain and occurs in the relatively small number of documents. In terms of CDC algorithm, a word has specific meaning if its context is of low entropy. Hence, we can define importance of a term in the following way:

$$imp(t) = \frac{1}{\log(1 + df(t))H(t)}. \tag{2}$$

One can see that the lower entropy of a term is, the higher its importance, and the bigger number of documents containing the term is, the lower its

importance. Also, since, for example, if a term occurs in 100 documents or in 101 makes smaller impact at the intuitive term importance than if a term occurs in 1 documents or in 2 document we applied *log* to document frequency of a term.

We use 2- or 3-word sequences in documents as phrases. The importance of a phrase is a sum of importance of terms it is composed of:

$$imp_1(p) = \sum_{t \in p} imp(t). \tag{3}$$

3 Experiments

We have made four experiments. In each experiment we use the same methodology with slight changes that allows us to compare results of our experiments. All our experiments are devoted to find out a response to the following question: how useful can be phrases and important words extracted from a document collection in automatic way using context information, and, particularly, what way should be chosen using word importance discrimination defined in (2) and (3). In the first experiment, $8T0eZ$, only phrases are used to score documents over topics. In the second experiment, $xLQOW$, we mix two types of scores obtained by a document against a topic: score obtained with common phrases in document and topic and score obtained with common important words. In the third experiment, $Krcy7$, we expand the list of phrases by phrases from documents retrieved in experiment $8T0eZ$. And the fourth experiment, $U2LwQ$, is the same as the first one but only "query" field is used.

3.1 Common part of experiments

The CSIRO document collection is parsed in the following way. First of all, we delete stop-words from the documents. In the experiments a word is defined as a string containing alphanumeric symbols and at least one letter, specifically a word satisfies the "[a-z0-9]*[a-z][a-z0-9]*" regular expression. Applying Porter stemming algorithm to words we obtain stem of words which we call terms. The number of terms we have got is 603349. A document is presented by profile which is a probability distribution defined by (1). We create contexts for term having document frequency greater or equal to 25 and less or equal to 10000. Term importance and phrase importance are calculated.

Parsing queries we concatenate the both "query" and "narr" fields to form a topic. We parse the topic deleting stop-words and applying stemming to words

defined by the same regular expression as for documents. Terms which are not in the dictionary of the collection are ignored. Each topic contains a lot of terms having different importance which should be estimated. In experiments we test several ways to estimate word importance.

3.2 8T0eZ

In the experiment, a document gains a score against a topic if the document has common phrases with the topic.

$$score_p(d|q) = \sum_{p \in d, p \in q} \left(imp_1 \left(p \right) * \log \left(pf \left(p|d \right) + 1 \right) \right), \tag{4}$$

where d is a document, p is a phrase, $imp_1(p)$ is phrase importance, $pf(p|d)$ is the number of occurrence of phrase p in document d, and the summation is done over common phrases of a topic q and document d. We report the first thousand documents for each topic having highest $score_p(d|q)$ values.

3.3 xLQOW

Sometimes the scores applied in experiment $8T0eZ$ are too strict and relaxing is required. In this experiment we mix document scores obtained with phrases and important words.

$$score_t(d|q) = \sum_{t \in d, t \in q} \left(imp(t) * \log(tf(t|d) + 1) \right), \tag{5}$$

where d is a document, t is a term, $imp(t)$ is term importance, $tf(t|d)$ is the number of occurrence of term t in document d, and the summation is done over common term of a topic q and document d.

We mix (4) and (5) scores:

$$score(d|q) = \lambda * score_p(d|q) + (1 - \lambda) * score_t(d|q),$$

where $0 < \lambda < 1$.

We optimized λ coefficient using topics and relevance judgements of TREC 2007. The optimal value of λ is 0.7.

We report the first thousand documents for each topic having highest $score(d|q)$ values.

3.4 Krcy7

In this experiment we use different kind of relaxing than in the previous one. We consider scores from experiment $8T0eZ$. Let us assume that we deal with topic q and document d which has a number of common phrases with topic q. The common phrases of document d and topic q are called phrases of level 1.

Let us define a set of documents containing given phrase p.

$$D(p) = \{d|p \in d\}.$$

We weight phrase p by its level 1 importance with scores of documents containing phrase p against all the topics:

$$imp_2(p) = \sum_{d \in D(p)} \sum_{q} (imp_1(p) * score_p(d|q)),$$

where q is a topic, d is a document, p is a phrase occurring in document d, $imp_1(p)$ is importance of phrase according (3). Hence, we get what we call "level 2 importance" of phrase. So, if a document has a phrase of level 1 against a topic it has a number of phrases of level 2 with level 2 importance. We use these phrases to expand list of phrases to search. We note that phrases of level 1 are phrases of level 2, too.

Let us define a set of documents having common phrases with a topic:

$$L(q) = \{d|\exists p, p \in d, p \in q\}.$$

and scoring function for documents against topics is

$$\begin{aligned} score_{p2}(d|q) = & \sum_{\substack{p \in d, \\ p \in q}} \left(imp_1\left(p\right) * \log\left(pf\left(p|d\right) + 1\right)\right) + \\ & + \sum_{\substack{p \in d, \\ d \in L(q)}} \left(imp_2\left(p\right) * \log\left(pf\left(p|d\right) + 1\right)\right). \end{aligned}$$

We report the first thousand documents for each topic having highest $score_{p2}(d|q)$ values.

3.5 U2LwQ

Documents are treated as in $8T0eZ$ experiment. The field "query" is used as a topic. We used all words in "query" field which were met in the collection. All

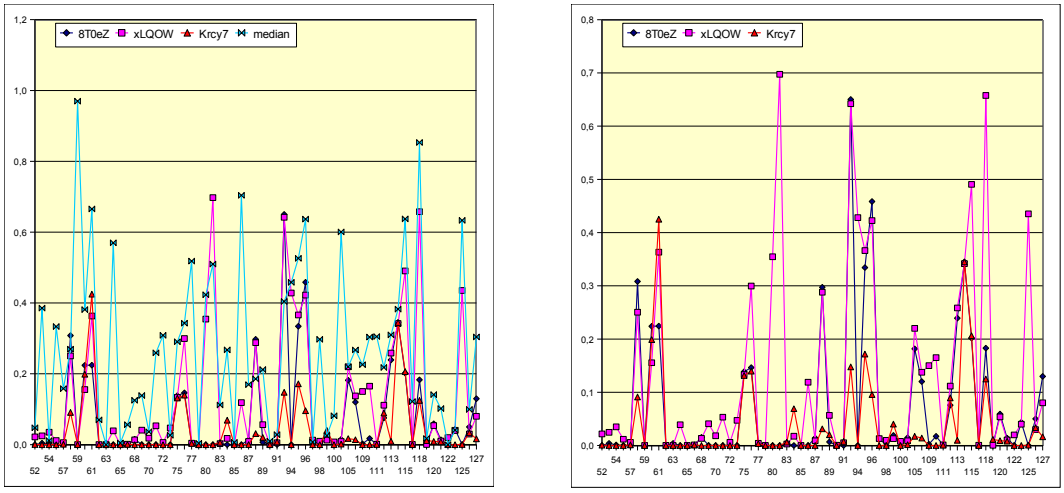

Fig. 1. AP measure on each topic with (left) and without (right) median results over all systems. Topics *id*s are placed at abscissa axis, and AP measure values are places at ordinate axis.

the other procedures are the same as in $8T0eZ$. We report the first thousand documents for each topic having highest $score(d|q)$ values.

4 Experimental results

The experimental results are presented in Table 1. The results confirm that the scores applied in $8T0eZ$ is too strict and the quality of retrieval can be improved by using important words equally with important phrases, as in experiment $xLQOW$. The attempt to use important phrases of level 2 does not give an advantage, see experiment $Krcy7$.

	$8T0eZ$	$xLQOW$	$Krcy7$	$U2LwQ$	*median*	*best*
infAP	0.0723	0.1300	0.0392	0.0339	0.2670	0.5541
infNDCG	0.1538	0.3057	0.1144	0.0742	0.4679	0.7803

Table 1
Average measures over all topics. *Median* is the average over of all the topics of the median measures of all the participated systems, and *best* is the average over of all the topics of the best achieved result among all the participated systems.

Experimental results at each topic are presented at Fig.1 and Fig.2. One can see from left graphs of Fig.1 and Fig.2 that results are worse than median results over all the systems for most topics but in two cases for AP measure and in three cases for NDCG measure results are better than median results. Observing right graphs of Fig.1 and Fig.2 we can see that relaxed scores applied in experiment $xLQOW$ performs better than scores of experiment $8T0eZ$ in most cases, but strict scores $8T0eZ$ and very relaxed scores $Krcy7$ can perform

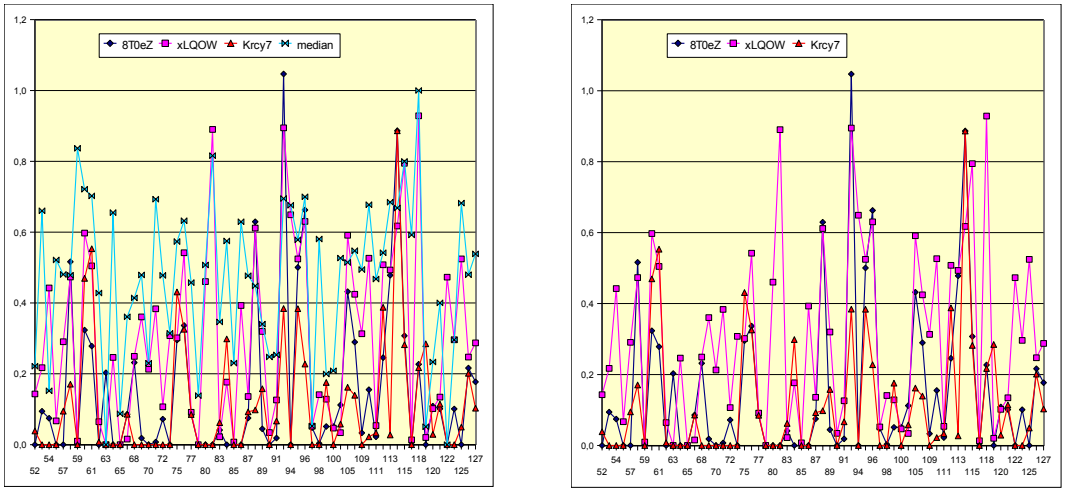

Fig. 2. NDCG measure on each topic with (left) and without (right) median results over all systems. Topics *id*s are placed at abscissa axis, and NDCG measure values are places at ordinate axis.

better than relaxed scores $xLQOW$ in some cases. Considering all the figures and content of topics we found out that better performance of scores, like over performing median results and achieving values of AP measure higher than 0.48 and NDCG measure higher than 0.78, is reached at short topics containing almost only informative important words.

References

[1] A. Bookstain, S.T. Klein, and T. Raita. Detecting content-bearing words by serial clustering - extended abstract. In *Sigir 1995*, 1995.

[2] Vladimir Dobrynin, David W. Patterson, and Niall Rooney. Contextual document clustering. In Sharon McDonald and John Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 167–180. Springer, 2004.

[3] H.P. Luhn. A statistical approach to the mechanized encoding and searching of literary information, 1957.

[4] Justin Picard. Finding content-bearing terms using term similarities. In *EACL*, pages 241–244, 1999.

[5] Niall Rooney, David W. Patterson, Mykola Galushka, and Vladimir Dobrynin. A scaleable document clustering approach for large document corpora. *Inf. Process. Manage.*, 42(5):1163–1175, 2006.

[6] Y. Takayama, R. Flournoy, S. Kaufmann, and S. Peters. Information retrieval based on domain-specific word associations. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING99)*, pages 155–161, 1999.