Informativeness for Adhoc IR Evaluation:
A measure that prevents assessing individual documents

Romain Deveaud[1], Véronique Moriceau[2], Josiane Mothe[3], and Eric SanJuan[1]

[1] LIA, Univ. Avignon, France,
romain.deveaud@gmail.com eric.sanjuan@univ-avignon.fr
[2] LIMSI-CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, France, moriceau@limsi.fr
[3] IRIT UMR5505, Univ. Toulouse, France, josiane.mothe@irit.fr

Abstract. Informativeness measures have been used in interactive information retrieval and automatic summarization evaluation. Indeed, as opposed to *adhoc* retrieval, these two tasks cannot rely on the Cranfield evaluation paradigm in which retrieved documents are compared to static query relevance document lists. In this paper, we explore the use of informativeness measures to evaluate *adhoc* task. The advantage of the proposed evaluation framework is that it does not rely on an exhaustive reference and can be used in a changing environment in which new documents occur, and for which relevance has not been assessed. We show that the correlation between the official system ranking and the informativeness measure is specifically high for most of the TREC *adhoc* tracks.

Keywords: Information retrieval, Evaluation, Informativeness, Adhoc retrieval

1 Introduction

Information Retrieval (IR) aims at retrieving the relevant information from a large volume of available documents. Evaluating IR implies to define evaluation frameworks. In *adhoc* retrieval, Cranfield framework is the prevailing framework [1]; it is composed of documents, queries, relevance assessments and measures. Moreover, document relevance is considered as independent from the document rank and generally as a Boolean function (a document is relevant or not to a given query) even though levels of relevance can be used [7]. Effectiveness measurement is based on comparing the retrieved documents with the reference list of relevant documents. Moreover, it is based on the assessment assumption, that is the relevance of documents is known in advance for each query. It implies that the collection is static since it is assessed by humans. Cranfield paradigm facilitates reproductibility of experiments: at any time it is possible to evaluate a new IR method and to compare it against previous results; this is one of its main strengths. However, such a framework is not usable in changing environments when new documents are continuously added.

As opposed to Cranfield document relevance independency assumption, informativeness expresses the dependency of document relevance and takes into

account the interactive nature of IR [8]. Indeed, one limitation of Cranfield-based evaluation is that relevance is encoded by documents [4]. Moreover, document relevance assessment is a clear limitation in dynamic context, when new documents are continuously added.

Nugget-based evaluation has been introduced to tackle this problem: rather than considering document relevance, it considers information relevance [4]. This method makes it possible to consider documents that have not been evaluated to be labeled as relevant or not, simply because they contain relevant information or not. Similar assumption is considered in automatic translation and automatic summarization evaluation. However this type of measure has not been intensively used in *adhoc* retrieval evaluation.

Our goal is to develop a method to evaluate *adhoc* IR using an informativeness measure to ensure reproductibility in dynamic document collections. To evaluate our method, we compare the system rankings we obtained using the informativeness measure proposed in [6] with the official system rankings based on document relevance, considering various TREC collections on *adhoc* tasks.

We show that the correlation between the official system rankings and the informativeness measure is specifically high for most of the TREC *adhoc* tracks.

The rest of the paper is organized as follows. Section 2 presents our evaluation framework which makes use of n-grams for informativeness-based evaluation applied to *adhoc* retrieval. We also present the *adhoc* retrieval collections we will be using. Section 3 presents and discusses the comparison of system rankings when using informativeness-based measures with the official ranking for the various *adhoc* retrieval collections/sub-tasks. Finally, Section 4 concludes this paper.

2 N-gram based measure for Adhoc Retrieval

The evaluation method we developed makes an informativeness measure being usable in the case of *adhoc* retrieval. We use it on various TREC adhoc tracks.

We use the generic Log Similarity (LogSim) informativeness measure initially introduced to evaluate tweet contextualization in 2011 CLEF-INEX QA task [5]. LogSim is based on pools of relevant passages extracted from the document collection (Wikipedia in the CLEF/INEX lab case) called t-rels. t-rels are chunks of texts that are marked-up as relevant by human assessors. By considering each n-gram word in these t-rels as a relevant item, the LogSim normalized measure is based on n-gram precision and graded using log frequencies.

Given a reference R and a summary S, the Log Similarity on n-grams ($LogSim$) measure stands as:

$$LogSim(S|R) = \sum_{w \in \mathcal{F}^n_{St}(R) \cap \mathcal{F}^n_{St}(S)} \frac{log(\min(P(w|S), P(w|R)).|R| + 1)}{log(\max(P(w|S), P(w|R)).|R| + 1)}.P(w|R) \quad (1)$$

where $P(w|X) = \frac{f_X(w)}{|X|}$ corresponds to the frequency $f_X(w)$ of n-gram w in X over the length $|X|$ of text X, $\cap\mathcal{F}^n_{St}(S(X)$ is the set of n-grams of stem words from X, and X is either R or S.

To build such textual references over document ad-hoc q-rels in order to easily apply informativeness to *adhoc* IR tasks, one approach consists in extracting from documents information nugget candidates; it has been shown that this is possible over non-spammed document collections like TREC robust track or Gov collections [4]. This paper aims at showing that similar results can be obtained without requiring a prior extraction of relevant nuggets. Indeed we propose a direct conversion of relevant documents into a textual reference and experiment plain informativeness measures over it.

For that, we introduce the concept of content interpolated precision at length λ (cP_λ). Assuming that a user reads the retrieved documents following the ranking given by an IR system, cP_λ evaluates the informativeness of the reading after λ words.

2.1 Evaluating Precision based on Document Content

Consider an *adhoc* task and its document q-rels. Assume that runs are ranked according to Mean Average Precision or Interpolated Precision at several recall levels. Runs can be converted into textual outcomes by concatenating ranked documents and q-rels can be converted into a textual reference by merging together all relevant documents per topic. Runs can be then evaluated by applying informativeness metrics to measure the overlap between submission and reference at various recall levels.

Let $D = (D_i)_{1 \leq i \leq d}$ be a ranked list of d documents. We consider as text $T = (t_1, ..., t_n)$ the concatenation of these documents (where $n = \sum_{i=1}^{i=d} |D_i|$). For each integer λ, we denote by T_λ the truncated text $T_\lambda = (t_1, ..., t_\lambda)$ and by $D_\lambda^{n,k}$ the set of n-grams with gap k. $D_\lambda^{n,0}$ or D_λ^n being the set of n-grams.

Similarly, given a set $R = \{R_i : 1 \leq i \leq r\}$ of r relevant documents we shall consider: $R^{n,k}$ the set of n-grams with gap k occurring in at least one reference.

In the case of TREC tracks, D is a run, each D_λ^n is the set of n-grams occurring in one of the m top ranked documents such that $\sum_{i=1}^{i=m} |D_i| \leq \lambda$ meanwhile R^n is the set of n-grams appearing at least once in the relevant documents from the corresponding q-rels.

We apply the English Porter stemming algorithm[4] to all documents after removing all stop words and all document identifiers like TREC doc-ID, etc. This is not only to reduce data, but to convert q-rels into reusable textual relevance judgments (t-rels) than can be applied to non official runs including documents not in the initial collection.

Given a run D and a reference R, we define for $n \in \{1, 2\}$, $k \in \{0, 2\}$ the content interpolated precision $cP_\lambda^{n,k}$ as:

$$cP_\lambda^{n,k}(D, R) = LS(D_\lambda^{n,k} | R^{n,k}) \tag{2}$$

Observe that, if $D_j \in R$ then :

$$cP^{1,0}_{\sum_{i=1}^{i=d} |D_i|}(D, R) \geq \frac{|D_j|}{\sum_{i=1}^{i=d} |D_i|}$$

[4] http://snowball.tartarus.org/algorithms/english/stemmer.html

but conversely, $D \cap R = \emptyset$ does not imply $cP_\lambda(D) = 0$ since there can be some overlap between n-grams in documents and in the reference.

So our approach based on document contents instead of document IDs does not require exhaustive references and therefore, can be applied to incomplete references based on pools of relevant documents. However, meanwhile *adhoc* IR returns a ranked list of documents independently of their respective lengths, relevance judgments can be used to automatically generate text references (*t-rels*) by concatenating the textual content of relevant documents.

2.2 Data Sets and Ground Truth

Among international evaluation collections, we chose TREC collections composed of news articles (Robust2004) and Web (Web and Terabyte). Doing so, we also focused on the quality of the collections with large amounts of runs and a comprehensive set of relevance judgments. The number of retrieval systems to rank ranges from 56 to 129, while the number of topics is typically 50 and increases to 150 for Terabyte2006 and 250 for Robust2004 (Table 1). All runs can be downloaded from the TREC web site, and document collections can be obtained on the web site for active participants or through track organizers.

Name	**# runs**	**# topics**	**Corpus**	$\lvert D_n^{1,1} \cup R^{1,1} \rvert$
TREC-5	106	50	TREC Vol. 4+5	15×10^6
TREC-6	107	50	TREC Vol. 4+5	12×10^6
TREC-7	103	50	TREC Vol. 4+5	28×10^6
TREC-8	129	50	TREC Vol. 4+5	35×10^6
Web2000	104	50	WT10g	137×10^6
Web2001	97	50	WT10g	195×10^6
Robust2004	110	250	TREC Vol. 4+5	150×10^6
Terabyte2004	70	50	GOV2	46×10^6
Terabyte2005	58	50	GOV2	46×10^6
Terabyte2006	80	150	GOV2	46×10^6
Web2010	56	50	ClueWeb09-B	66×10^6
Web2011	62	50	ClueWeb09-B	64×10^6

Table 1: **Summary of TREC test collections and size in tokens of generated t-rels used for evaluation.**

We take the same experimental approach as in [3] and [2], and reproduced the official rankings of all these retrieval systems for these various collections using the official measure. For all collections, the official measure is the Mean Average Precision (MAP), except for Web2010 and Web2011 where Expected Reciprocal Rank (ERR@20) was preferred. These official rankings constitute the ground truth ranking, against which we will compare the rankings produced by:

– cP_{10^3} based on the 1000 tokens of each run and on their log frequencies.

– cP_n based on all tokens of each run.

By comparing averaged measures, we evaluate if the average informativeness of all documents retrieved by a system is correlated to the official ranking. Let us emphasize that this does not necessarily imply that document informativeness is correlated to individual document relevance for a given query. We use the Kendall's τ rank correlation coefficient to identify correlations between the ground truth ranking and the informativeness ranking.

3 Results and discussion

In this section we report the correlation results of the ground truth ranking (TREC official measure depending on the track) and the content-based ranking produced by the cP_λ informativeness measure. All correlations reported are significantly different from zero with a p-value < 0.001. While we chose Kendall's τ as the correlation measure, we also report the Pearson's linear correlation coefficient for convenience. A $\tau > 0.5$ typically indicates a strong correlation since it implies an agreement between the two measures over more than half of all ordered pairs.

Track	$cP^{1,1}_{10^3}$		$cP^{2,0}_{10^3}$		$cP^{2,2}_{10^3}$		$cP^{1,1}_{n}$	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ
TREC-5	57.91%	71.04%	56.22%	56.30%	56.15%	56.25%	**69.91%**	**88.62%**
TREC-6	72.49%	84.18%	76.50%	92.90%	**76.50%**	**93.01%**	58.78%	68.18%
TREC-7	61.27%	82.17%	70.18%	90.59%	**70.27%**	**90.60%**	63.45%	53.49%
TREC-8	54.80%	84.04%	65.79%	92.25%	65.85%	**92.30%**	**67.46%**	72.94%
Web2000	46.48%	68.00%	61.42%	83.40%	62.05%	77.82%	**70.83%**	**86.68%**
Web2001	31.65%	56.51%	36.94%	57.89%	36.66%	56.66%	**77.45%**	**88.90%**
Robust2004	40.97%	64.33%	58.67%	85.43%	59.50%	86.22%	**74.71%**	**90.88%**
Terabyte2004	48.56%	60.12%	61.25%	73.65%	61.45%	74.75%	**76.37%**	**86.12%**
Terabyte2005	59.45%	85.34%	69.65%	88.98%	69.80%	**89.18%**	**76.01%**	84.28%
Terabyte2006	41.14%	50.24%	54.75%	70.70%	55.28%	70.90%	**65.04%**	**89.37%**
Web	28.56%	44.00%	**44.38%**	**69.11%**	44.17%	69.09%	-	-
Web2011	56.03%	**80.98%**	55.68%	79.14%	**56.35%**	79.42%	**34.50%**	-

Table 2: **Retrieval systems ranking correlations between the official ground truth and the cP_λ informativeness measure. $cP^{1,1}_\lambda$ stands for uniterms while $cP^{2,2}_\lambda$ corresponds to bigrams with skip. We use either 10^3 terms or all the terms from the ordered list of retrieved documents.**

When looking at Table 2, we see that cP_λ accurately reproduces official ranking based on MAP for early TREC tracks (TREC6-7-8, Web2000) as well as for Robust2004, Terabyte2004-5 and Web2011. *LogSim*-score applied to all tokens in runs is often the most effective whenever systems are ranked based on MAP.

However on early TREC tracks, $cP^{2,2}_{10^3}$ can perform better even though only the first 1000 tokens of each run are considered after concatenating ranked retrieved documents. Indeed, the traditional TREC *adhoc* and Robust tracks used newspaper articles as document collection. Since a single article often deals with a single subject, relevant concepts are likely to occur together, which might be less the case in web pages for example. A relevant news article is very likely to contain only relevant information, whereas a long web document that deals with several subjects might not be relevant as a whole.

4 Conclusion

In this paper, we proposed a framework to evaluate adhoc IR using the LogSim informativeness measure based on token n-grams. To evaluate this measure, we compared the ranks of the systems we obtained with the official rankings based on document relevance, considering various TREC collections on *adhoc* tasks. We showed that 1) rankings obtained based on n-gram informativeness and with Mean Average Precision are strongly correlated; and 2) $LogSim$ informativeness can be estimated on top ranked documents in a robust way. The advantage of this evaluation framework is that it does not rely on an exhaustive reference and can be used in a changing environment in which new documents occur, and for which relevance has not been assessed. In future work, we will evaluate various $LogSim$ parameters influence.

References

1. C. Cleverdon. The Cranfield tests on index languages devices. 1967.
2. C. Hauff. Predicting the effectiveness of queries and retrieval systems. PhD thesis, Enschede, January 2010. SIKS Dissertation Series No. 2010-05.
3. R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. Inf. Process. Manage., 42(3):595–614, May 2006.
4. V. Pavlu, S. Rajput, P. B. Golbus, and J. A. Aslam. IR System Evaluation Using Nugget-based Test Collections. In Proceedings of WSDM, 2012.
5. E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2011 Question Answering track (QA@INEX). In Focused Retrieval of Content and Structure. Springer, 2012.
6. E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2012 Tweet Contextualization Track. In CLEF, 2012.
7. A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. Information Processing and Management, 34(5):599 – 621, 1998.
8. J. Tague-Sutcliffe. Measuring the informativeness of a retrieval process. In Proceedings of the International ACM SIGIR conference on research and development in Information Retrieval, pages 23–36, 1992.