Information Processing and Management 51 (2015) 791–808

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Evaluating document filtering systems over time

Tom Kenter[a,*], Krisztian Balog[b], Maarten de Rijke[a]

[a] University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[b] University of Stavanger, Stavanger, Norway

ARTICLE INFO

Article history:
Received 22 September 2014
Revised 8 February 2015
Accepted 27 March 2015
Available online 8 September 2015

Keywords:
Time-aware information retrieval
Evaluation
Significance testing

ABSTRACT

Document filtering is a popular task in information retrieval. A stream of documents arriving over time is filtered for documents relevant to a set of topics. The distinguishing feature of document filtering is the temporal aspect introduced by the stream of documents. Document filtering systems, up to now, have been evaluated in terms of traditional metrics like (micro- or macro-averaged) precision, recall, MAP, nDCG, F1 and utility. We argue that these metrics do not capture all relevant aspects of the systems being evaluated. In particular, they lack support for the temporal dimension of the task. We propose a time-sensitive way of measuring performance of document filtering systems over time by employing trend estimation. In short, the performance is calculated for batches, a trend line is fitted to the results, and the estimated performance of systems at the end of the evaluation period is used to compare systems. We detail the application of our proposed trend estimation framework and examine the assumptions that need to hold for valid significance testing. Additionally, we analyze the requirements a document filtering metric has to meet and show that traditional macro-averaged true-positive-based metrics, like precision, recall and utility fail to capture essential information when applied in a batch setting. In particular, false positives returned in a batch for topics that are absent from the ground truth in that batch go unnoticed. This is a serious flaw as overgeneration of a system might be overlooked this way. We propose a new metric, aptness, that does capture false positives. We incorporate this metric in an overall score and show that this new score does meet all requirements. To demonstrate the results of our proposed evaluation methodology, we analyze the runs submitted to the two most recent editions of a document filtering evaluation campaign. We re-evaluate the runs submitted to the Cumulative Citation Recommendation task of the 2012 and 2013 editions of the TREC Knowledge Base Acceleration track, and show that important new insights emerge.

1. Introduction

Document filtering is a popular task in information retrieval with many applications (Keiser, 2009; Amigó, Gonzalo, & Verdejo, 2011; Amigó, Gonzalo, & Verdejo, 2013; Robertson & Soboroff, 2002; Frank, Kleiman-Weiner, et al., 2013; Frank, Bauer, et al., 2013). A stream of documents arriving over time is filtered for documents relevant to a set of topics. The distinguishing feature of document filtering, that sets it apart from other document classification tasks, is the temporal aspect introduced by the stream of documents. Because of this temporal dimension, the performance of a system is

* Corresponding author.
E-mail addresses: tom.kenter@uva.nl (T. Kenter), krisztian.balog@uis.no (K. Balog), derijke@uva.nl (M. de Rijke).

http://dx.doi.org/10.1016/j.ipm.2015.03.005
0306-4573/

susceptible to change over time. For example, in a document filtering setting, where topics are being monitored over time, the topics might evolve. A system filtering the stream should be sensitive to this in order to perform well. As another example, a spam classifier should adapt to malignant and cunning adversaries. If it fails to do so effectively, its performance will likely degrade over time.

Standard evaluation metrics measure performance for all returned documents, for a given information need (i.e., query), in a single batch (e.g., P@n, recall, nDCG, AP) and they can be averaged over multiple requests (e.g., macro-precision across a set of queries). Recent studies propose to decompose document streams into sequences (e.g., into slices of equal size) and measure effectiveness on a given time period (Azzopardi, 2009; Dietz, Dalton, & Balog, 2013; Aslam, Ekstrand-Abueg, Pavlu, Diaz, & Sakai, 2013). Then, it becomes possible to monitor the changes in system performance over time and to measure performance as a weighted average of slice-based relevance scores. None of these approaches, however, addresses the question we are interested in: how system performance *changes* over time.

In Fig. 1 the performance of three hypothetical systems is plotted over time. The blue dots represent the score of the system at a given point in time. The gray dotted line represents the average performance of the systems over the entire time span. Clearly, all three systems have the same average performance. However, the performance of System A degrades rather strongly over time, the performance of System B less so, while the performance of system C shows improvement over time. With the metrics currently available there is no way to express this difference. In this paper we propose to capture this difference by employing trend analysis. In short, this entails fitting a straight line to the values of any existing performance metric applied to document filtering systems over time. In Fig. 1 these lines are displayed in orange. The derivative of the fitted line provides a simple and intuitive measure of the amount of change in performance over time. Ultimately, we can compare the performance of the three systems as estimated by trend analysis at the end of the evaluation period (the large orange dots in Fig. 1).

The main contributions of this paper are the following. We analyze the properties and requirements a document filtering metric should meet. We propose to measure performance in batches and to use trend estimation to measure performance over time. We show that traditional macro-averaged true-positive-based metrics, like precision, recall and F1 fail to capture essential information when applied in a batch setting. In particular, documents returned in a batch for topics that are absent from the ground truth in that batch, false positives, go unnoticed. We propose a new metric, aptness, that does capture false positives. We incorporate this metric in an overall score, F_{pra}, and show that this new score does meet all requirements. As an important aspect of evaluation is testing for significant differences in observations, we detail the tests for statistical significance for trend estimation and discuss the assumption that need to hold.

We test our method on the runs submitted to the Cumulative Citation Recommendation task of the 2012 and 2013 editions of the TREC Knowledge Base Acceleration track (*KBA CCR* for short). A re-evaluation of the results in terms of our proposed method shows a different ordering of teams, also at the top end. Moreover, while there were teams beating the baseline in 2013 when judged by the official metrics, our tests show that in fact no team did, when our proposed time-aware evaluation is used.

Additionally, we find that the assumptions needed for valid significance testing hold in a vast majority of cases considered.

The remainder of the paper is organized as follows. Related research is covered in Section 2. In Section 3 we give an overview our time-aware document filtering evaluation method and discuss the required properties. Two key components of our approach, measuring performance per time batches and trend estimation are then presented in Sections 4 and 5, respectively. In Section 5 we also show how significance testing can be performed. In Section 6 we detail our research questions and show results of experiments performed on the runs submitted to two years of the TREC KBA CCR evaluation campaign. We conclude in Section 7.

2. Related work

2.1. Document filtering

Document filtering is the task of identifying relevant items from an incoming stream of content. Different flavors of this problem have been studied in the past. Common to these is that documents arrive sequentially over time and relevance decisions for each topic must be made as soon as the document is processed. Topics represent long-term information needs (often referred to as "profiles") and may evolve over time.

Routing was one of the very first tasks studied at the Text REtrieval Conference (TREC) and ran at the first three editions of TREC (TREC-1-3) (Harman, 1994). Routing systems learn static profiles from training documents, and then rank documents in the test set according to these profiles. Consequently, performance evaluation relies on ranked-based measures. "After TREC-3, a strong argument was made that a more realistic filtering task should be developed in addition to routing." (Soboroff & Robertson, 2003). Early editions of the Filtering track (TREC-4-6) cast the task as a binary classification problem: to refer the document to the user or not. This necessitates a methodological departure from rank-based evaluation to the use of set-based measures (Soboroff & Robertson, 2003). Later editions (from TREC-7) further refine filtering into *batch filtering* and *adaptive filtering* tasks. Systems have to process test documents in chronological order and select a subset of them. This implies the use of thresholding for making the binary decision between selecting or discarding each document. Systems are