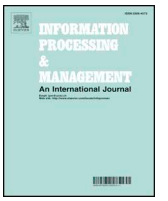
Information Processing and Management 57 (2020) 102149

Contents lists available at ScienceDirect

Information Processing and Management

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



Effectiveness evaluation without human relevance judgments: A systematic analysis of existing methods and of their combinations

Kevin Roitero, Andrea Brunello, Giuseppe Serra, Stefano Mizzaro*

University of Udine, Udine, Italy

ARTICLE INFO

Keywords:
Information retrieval evaluation
Automatic evaluation
Machine learning
Topic difficulty

ABSTRACT

In test collection based evaluation of retrieval effectiveness, it has been suggested to completely avoid using human relevance judgments. Although several methods have been proposed, their accuracy is still limited. In this paper we present two overall contributions. First, we provide a systematic comparison of all the most widely adopted previous approaches on a large set of 14 TREC collections. We aim at analyzing the methods in a homogeneous and complete way, in terms of the accuracy measures used as well as in terms of the datasets selected, showing that considerably different results may be achieved considering different methods, datasets, and measures. Second, we study the combination of such methods, which, to the best of our knowledge, has not been investigated so far. Our experimental results show that simple combination strategies based on data fusion techniques are usually not effective and even harmful. However, some more sophisticated solutions, based on machine learning, are indeed effective and often outperform all individual methods. Moreover, they are more stable, as they show a smaller variation across datasets. Our results have the practical implication that, when trying to automatically evaluate retrieval effectiveness, researchers should not use a single method, but a (machine-learning based) combination of them.

1. Introduction

In Information Retrieval (IR), test-collection based effectiveness evaluation is a well-known and quite standard method. The whole evaluation process has a cost, in terms of resources needed, effort made by the research community, and also money; thus it is not surprising that researchers tried and are still trying to reduce such costs, for example by using fewer topics, more sensitive effectiveness metrics, shallower pools, or cheaper (usually, crowdsourced) human relevance judgments. A more radical approach is to avoid human relevance judgments altogether, as it has been proposed by several researchers (Aslam & Savell, 2003; Diaz, 2007; Nuray & Can, 2003; 2006; Sakai & Lin, 2010; Soboroff, Nicholas, & Cahan, 2001; Spoerri, 2007; Wu & Crestani, 2003). In this paper, we set out to provide a detailed and complete analysis of the methods for effectiveness evaluation without human relevance judgments, as well as study if they can be fruitfully combined. More in detail, we review the methods that have been proposed (Section 2), we outline the motivations for this work and propose three research questions (Section 3), we describe the experimental setting (Section 4), and we answer each research question (Sections 5–7).

* Corresponding author.
E-mail addresses: kevin.roitero@spes.uniud.it (K. Roitero), andrea.brunello@uniud.it (A. Brunello), giuseppe.serra@uniud.it (G. Serra), mizzaro@uniud.it (S. Mizzaro).

https://doi.org/10.1016/j.ipm.2019.102149
Received 1 March 2019; Received in revised form 27 August 2019; Accepted 20 October 2019

2. Related work

Table 1 summarizes the proposals to use no human relevance assessments when evaluating IR effectiveness. The first proposal is by Soboroff et al. (2001): Their method performs a random sample from the pool of documents (i.e., the documents retrieved by at least one system); the sampled documents are deemed to be relevant, while the remaining ones are non relevant, and the evaluation is performed accordingly. The underlying assumption is that relevant documents tend to be retrieved by many systems, and thus pooled.

Another method, proposed by Wu and Crestani (2003), is based on data fusion, and consists in merging the ranked lists of documents retrieved by each retrieval system querying the same test collection for a certain topic. The idea is to assign a weight to each retrieved document and to use such weights to rank retrieval systems. Thus, good systems are those that retrieve "popular" documents. In the simplest version of the algorithm (WUCv0), the weight, called reference count, sums up the occurrences of each document retrieved by a system which is present in the ranked lists of other systems. The four variants assign a weight to the reference count differentiating the position in which each document appears in the ranked list.

Aslam and Savell's method (2003) measures the similarity of each system to the others (by computing the ratio between the cardinality of the intersection of the documents of the ranked lists and their union) and uses this similarity to evaluate them. This evaluation is highly correlated to Soboroff et al.'s method. One issue is that the average similarity is computed by means of "the grossest possible measure" (Aslam & Savell, 2003, p. 362). This work also presents one of the main criticisms to this approach: The observation that runs are ranked by popularity rather than effectiveness. Such "tyranny of the masses" effect is penalizing for best runs, that are underestimated. We use a slightly modified version of this method, keeping the raw topic scores instead of computing their mean value over the topic set.

The method by Nuray and Can (2003, 2006) consists of three phases: (i) select the runs, (ii) compute the popularity of each document according to various methods, and (iii) the top 30% of the most popular documents are said to be relevant. The run selection can be done in two ways: either "normal", where each run is selected, or "bias", where the runs selected are the top 50% of runs which have a list of retrieved document that is farther from the "norm". The document ranking can be performed according to three strategies taken from theory of voting: "Rank Position", "Borda" (Emerson, 2013), and "Condorcet" (Fishburn, 1977).

The method by Spoerri (2005) selects one run for each participating organization, and forms a set of trials containing five runs (we borrow this terminology from Sakai & Lin, 2010) in a way that each run occurs exactly five times (in different trials); then, it computes the percentage of the set of documents either retrieved by the run exclusively (called "Single"), the set of documents retrieved by all the five runs in the trial ("AllFive"), and the "Single minus AllFive" measure. Finally, to obtain a trial-independent behavior, the three computed measures for each run are averaged over the five trials in which the run occurs.

The method by Sakai and Lin (2010) is very similar to Condorcet method, even if statistically different and more efficient.

All the above methods have been experimentally evaluated using as datasets some TREC test collections as detailed in Table 1

Table 1
The 17 prediction methods used in this paper.

#	Acronym (version)	Accuracy Measures	Datasets	Effectiveness Metrics
Soboroff et al. (2001)		τ, charts	TREC 3,5,6,7,8	MAP
1	SNC			
Wu and Crestani (2003)		r_s	TREC 3,5,6,7,	R-Precision,
	2001	P@10,30,50,100		
2	WUCv0 (Basic)			
3	WUCv1 (Variation 1)			
4	WUCv2 (Variation 2)			
5	WUCv3 (Variation 3)			
6	WUCv4 (Variation 4)			
Aslam and Savell (2003)		τ, ρ,	TREC 3,5,6,7,8	MAP
scatterplots				
7	AS			
Nuray and Can (2006)		r_s	TREC 3,5,6,7	MAP
8	NC-NRP (Normal Rank Position)			
9	NC-NB (Normal Borda)			
10	NC-NC (Normal Condorcet)			
11	NC-BRP (Bias Rank Position)			
12	NC-BB (Bias Borda)			
13	NC-BC (Bias Condorcet)			
Spoerri (2007)		ρ, scatterplots	TREC 3,6,7,8	MAP, P@1000
14	SPO-S (Single)			
15	SPO-A (AllFive)			
16	SPO-SA (Single - AllFive)			
Sakai and Lin (2010)		τ, τ_{ap}, charts, scatterplots	R03, R04, CLIR6-JA, CLIR6-CT, IR4QA-CS	MAP, nDCG, Q-measure
17	SL			