Do Easy Topics Predict Effectiveness Better Than Difficult Topics?

Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro(✉)

Department of Mathematics, Computer Science, and Physics, University of Udine,
Udine, Italy
roitero.kevin@spes.uniud.it, {eddy.maddalena,mizzaro}@uniud.it

Abstract. After a network-based analysis of TREC results, Mizzaro and Robertson [4] found the rather unpleasant result that topic ease (i.e., the average effectiveness of the participating systems, measured with average precision) correlates with the ability of topics to predict system effectiveness (defined as topic hubness). We address this issue by: (i) performing a more detailed analysis, and (ii) using three different datasets. Our results are threefold. First, we confirm that the original result is indeed correct and general across datasets. Second, we show that, however, that result is less worrying than what might seem at first glance, since it depends on considering the least effective systems in the analysis. In other terms, easy topics discriminate most and least effective systems, but when focussing on the most effective systems only this is no longer true. Third, we also clarify what happens when using the GMAP metric.

1 Introduction

Effectiveness evaluation of Information Retrieval (IR) systems is performed within many initiatives, such as TREC, NTCIR, INEX, CLEF, and others. Participants to these initiatives can test their own retrieval system over a set of *topics*, which are a representation of information needs. The effectiveness of each system is assessed considering various metrics such as Average Precision (AP) and Mean AP (MAP), that determine the final rank of systems. Interactions between the topics and the systems, and in particular between the difficulty of the topic and the final rank of the systems have been studied by Mizzaro and Robertson [4] considering link analysis techniques, and in particular the HITS algorithm [2]. More in detail, Mizzaro and Robertson investigate the correlation between topic ease and the ability to predict system effectiveness. They find that easier topics are better at estimating system effectiveness. In other terms, to be effective in TREC a system has to perform well on easy topics. This is undesirable since it is the difficult topics that are more interesting and it is by working on them that the state of the art of the discipline can advance most. In this paper we extend their analysis.

J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 605–611, 2017.
DOI: 10.1007/978-3-319-56608-5_55

2 Background

In an attempt to make this paper self contained, in this section we summarize the methodology and the relevant results of Mizzaro and Robertson [4] (for further details see the original paper). The output of the TREC competition can be represented as a table having as column the topics and as rows the systems (Fig. 1(b)). Each cell contains an effectiveness measure of each system on each topic. Effectiveness is measured according to some metric, and Average Precision (AP) is a common choice. In order to provide the final rank of systems, each row (i.e., in TREC terms, a *run*) is averaged to compute the MAP metric. Mizzaro and Robertson [4] introduce a dual metric, named Average AP (AAP), that is computed averaging over each column and represents the average value over the systems of the APs values for each topic: while MAP measures system effectiveness, AAP measures topic ease. The topic-system matrix of Fig. 1(b) is then normalized by transforming each AP value into $\overline{\mathrm{AP}_{\mathrm{A}}}(s_i, t_j)$ (Normalized AP according to AAP) and $\overline{\mathrm{AP}_{\mathrm{M}}}(s_i, t_j)$ (Normalized AP according to MAP):

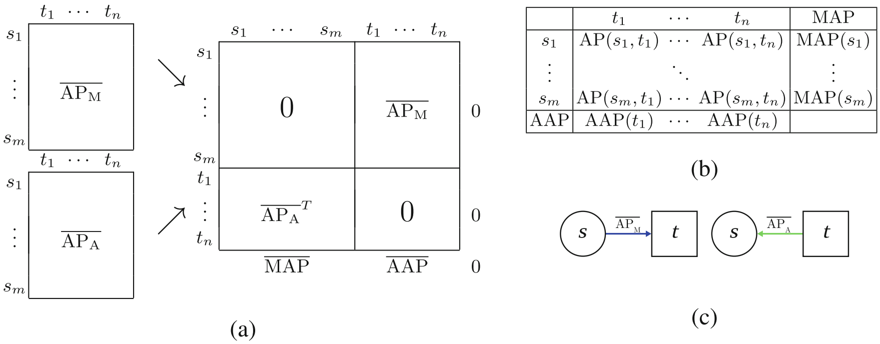
$$\overline{\mathrm{AP}_{\mathrm{A}}}(s_i, t_j) = \mathrm{AP}(s_i, t_j) - \mathrm{AAP}(t_j) \text{ and}$$
$$\overline{\mathrm{AP}_{\mathrm{M}}}(s_i, t_j) = \mathrm{AP}(s_i, t_j) - \mathrm{MAP}(s_i).$$

The two matrices obtained from the normalization process (one obtained considering $\overline{\mathrm{AP}_{\mathrm{A}}}$ and one from $\overline{\mathrm{AP}_{\mathrm{M}}}$) are used to study interactions between systems and topics; this step is accomplished by building an adjacency matrix, and, consequently, the corresponding graph, made of the two normalized matrices; this process is summarized in Fig. 1(a). Each link between the system and the topic (see Fig. 1(c)) represents [4]:

- arc $s \rightarrow t$ with weight $\overline{\mathrm{AP}_{\mathrm{M}}}$: how much the system s "thinks" that the topic t is easy (or "un-easy" if the weight is negative);
- arc $s \leftarrow t$ with weight $\overline{\mathrm{AP}_{\mathrm{A}}}$: how much the topic t "thinks" that the system s is effective (or "un-effective" if the weight is negative).

The graph is used to compute *hubness* and *authority*, obtained using an extended version of the HITS algorithm [2] which allows to include the negative values for the arcs. The authority of a topic measures topic ease, while the authority of a system measures system effectiveness [4]. The hubness of a topic measures the topic capability to recognize effective systems, while the hubness of a system measures its ability to recognize easy topics [4]. When focussing on the values of AAP and topic hubness, as we do in our paper, Mizzaro and Robertson [4, pp. 483–484] state:

> "[...] easier topics are better at estimating system effectiveness. [the statement] is a bit worrying. It means that system effectiveness in TREC is affected more by easy topics than by difficult topics, which is rather undesirable for quite obvious reasons: a system capable of performing well on a difficult topic, i.e., on a topic on which the other systems perform badly,


$t_1 \cdots t_n$
s_1
s_m
$\overline{\mathrm{AP}_\mathrm{M}}$
$\overline{\mathrm{AP}_\mathrm{A}}$
$s_1 \cdots s_m \; t_1 \cdots t_n$
0
$\overline{\mathrm{AP}_\mathrm{A}}^T$
t_1
t_n
$\overline{\mathrm{MAP}}$
$\overline{\mathrm{AAP}}$
0
0
0
(a)
(b)
s
t
$\overline{\mathrm{AP}_\mathrm{M}}$
$\overline{\mathrm{AP}_\mathrm{A}}$
(c)

	t_1	$\cdots$	t_n	MAP
s_1	AP(s_1, t_1)	$\cdots$	AP(s_1, t_n)	MAP(s_1)
$\vdots$		$\ddots$		$\vdots$
s_m	AP(s_m, t_1)	$\cdots$	AP(s_m, t_n)	MAP(s_m)
AAP	AAP(t_1)	$\cdots$	AAP(t_n)	

Fig. 1. (a) Construction of the adjacency matrix. $\overline{\mathrm{AP}_\mathrm{A}}^T$ is the transpose of $\overline{\mathrm{AP}_\mathrm{A}}$. (b) AP, MAP and AAP. (c) The relationships between systems and topics (from [4]).

> would be an important result for IR effectiveness; conversely, a system capable of performing well on easy topics is just a confirmation of the state of the art."

This statement is obtained when commenting Fig. 5(d) in [4] (we present a slightly modified version of that figure in Fig. 2(g), analysed in more detail the following). It is also noted that the correlation between AAP and hubness disappears when using GMAP (Geometric MAP [5]) in place of MAP (and GAAP in place of AAP) [4, p. 484]:

> "with GMAP[...] and GAAP [...] the correlation with hubness largely disappears"

3 Experiments

Aims and Settings. In this paper we further study the above results, analysing whether they hold if different subsets of systems are considered. We perform the same analyses, but we also repeat them using a subset of systems: we rank the systems according to their effectiveness (measured using MAP) and we select either the most or the least effective ones. More in detail, our procedure can be described as:

```
for cardinality n in range 1 to number of systems:
    order the systems according to MAP;
    select the first/last n systems;
    build the adjacency matrix;
    compute hubness (and authority) using HITS;
    compute Pearson's correlation between hubness and AAP;
```

Also, [4] used a single dataset; to study the generality of the results, in our experiments we consider the following three datasets:

– AH99: Ad-Hoc track, TREC 1998; this track has 129 systems and 50 topics. This is the dataset used in [4].
– TB06: Terabyte track, TREC 2006; this track has 61 systems and 149 topics.
– R04: Robust track, TREC 2004; this track has 110 systems and 249 topics.

Results. Figures 2 from (a) to (c) show the first results, one figure for each dataset. The x-axis shows the cardinality: at cardinality n we have considered the n best (most effective) or worst (least effective) systems; the maximum value on the axis is the number of systems participating in the TREC track. The y-axis shows the Pearson's correlation between hubness and AAP. In each figure the two blue series represent, respectively: the "Worst2Best" series (in the upper part of each figure) shows the correlation values when considering the systems ranked in ascending order of MAP (i.e., from least effective systems to most effective ones), and the "Best2Worst" series shows the correlation values when considering the systems ranked in descending order of MAP (i.e., from most effective systems to least effective ones). The red series are similar, with GMAP and GAAP in place of MAP and AAP.

Let us focus first on the top-right of the charts, i.e., on the maximum cardinality, where all the systems are considered (and the two series, of course, meet). That is the correlation value studied by Mizzaro and Robertson [4], and it is indeed high (a bit smaller for R04), substantially confirming previous results. The figures show something more, though, and in a consistent way over the three datasets. When considering the "Worst2Best" series (i.e., the systems ranked from the lowest to the highest MAP values), and moving towards left in the charts, the correlation values remain high when decreasing cardinality down to around 25% of the total systems, before decreasing and becoming noisy. This would still confirm the undesired effect, also taking into account that the noisy behaviour at low cardinalities can depend on the very low MAP of systems. However, when considering the "Best2Worst" series (i.e., the systems sorted from the highest to the lowest MAP values), the behaviour is different, and again similar for the three datasets: starting from low cardinalities, and moving towards the right in the charts, the correlations remain stable at near-zero values (let us say, in $[-0.5, 0.5]$) until about 90% of the maximum cardinality and then the correlation increases to the value obtained considering all the systems. Summarizing, the "undesired" feature that system effectiveness is affected mostly by easy topics manifests itself, with correlation values that are clearly different from zero, if and only if the very least effective systems (i.e., the bottom 10–25% or so) are considered, as shown by the "Best2Worst" series. The undesired feature is caused mostly by the least effective systems.

The scatterplots in Figs. 2(d)–(g) show the details of the AAP-hubness correlation for selected cardinalities (20, 100, 120, and 129, highlighted in Fig. 2(a) with the larger white dots) for the "Best2Worst" series of the AH99 dataset (other datasets are similar). The charts confirm that there is no correlation up to cardinality 100; some correlation exists at cardinality 120 and at full cardinality 129 we obtain the same result as [4, Fig. 5(d)]. Figure 2(h) shows the AAP-hubness correlation again at cardinality 20 and for the same dataset, but

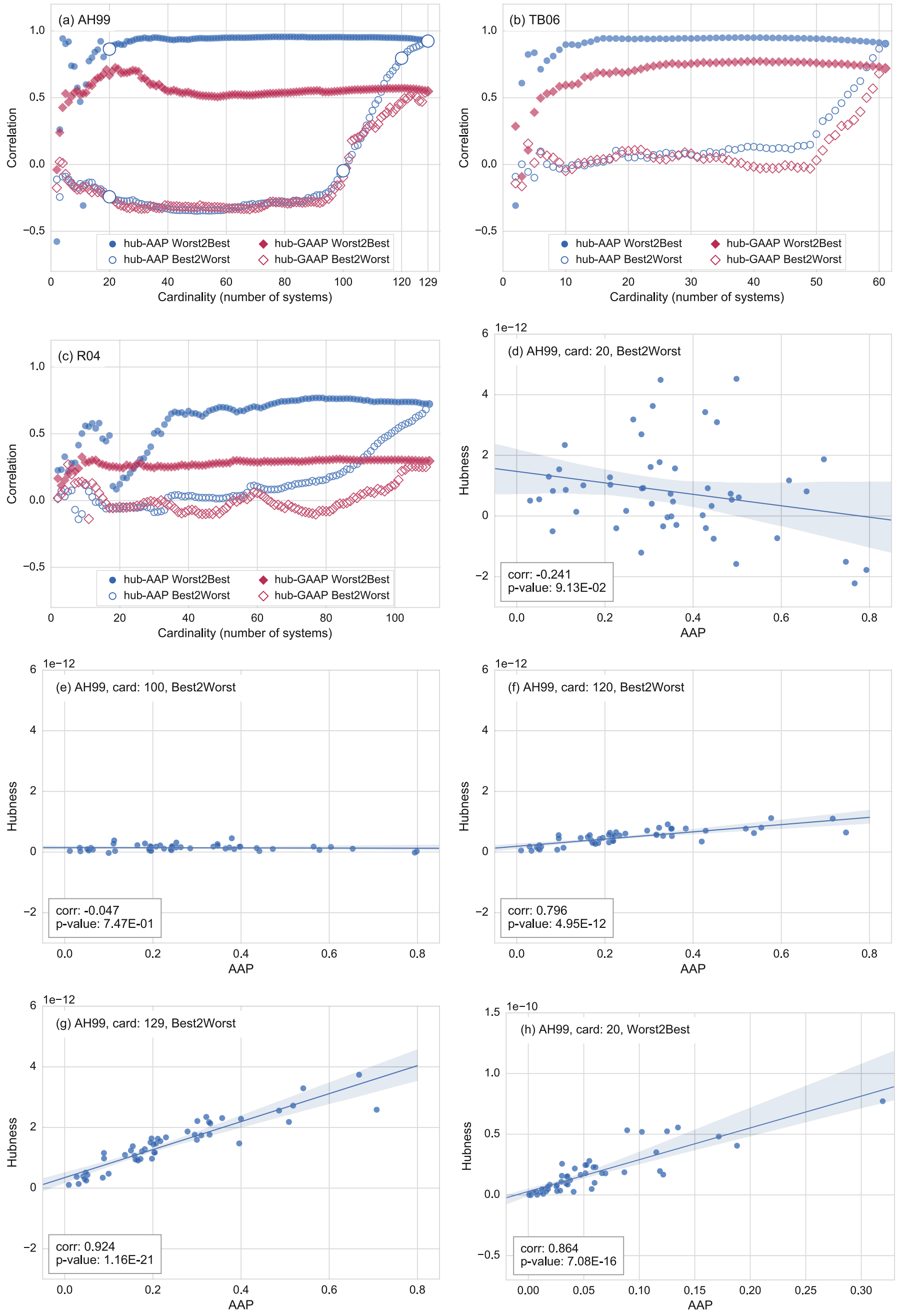


Fig. 2. Correlations for different systems subsets and metrics (a)–(c). Scatterplots for specific systems subsets at different cardinality values (d)–(h).

for the "Worst2Best" series, confirming that when considering only the worst systems the correlation appears much earlier, at lower cardinality.

We also repeated the same experiments using the GMAP (and GAAP) measure instead of MAP (and AAP). Differently from [4] in which the correlation largely disappears when using GMAP, our results show that some correlation still occurs (even if the GAAP-hubness correlation is much lower than the AAP-hubness one, due to the definition of GMAP that weighs less the easy topics), and the trend of the red series in the plots is comparable to that of the blue series, for both the "Best2Worst" and "Worst2Best" series. The effect of the least effective systems is still clear, even if smaller, also with GMAP and GAAP.

Although the general trends are the same across the three datasets, there are some differences. As mentioned above, correlation values are smaller for R04. Again for the same dataset, the growth of the correlation values is less sudden for the "Best2Worst" series. Both these results can depend on the peculiar features of R04: since it contains the most difficult topics from previous TREC editions, the systems have similar (and in general low) AP values, the variance of AP will be smaller, and this in turn might cause a lower correlation. Another difference is that correlation values for GAAP are much higher in the TB06 dataset: the benefical effect of GMAP reported in [4], besides being weaker in our experiments, almost disappears for this dataset. This requires further study.

4 Conclusions and Future Work

This paper presents an analysis that exploits some of the hidden relations between topics and systems used in TREC-like competitions. We have obtained three results: (i) we confirmed the original results that easier topics are better in distinguishing system effectiveness, and generalized it to different datasets; (ii) however, we also somehow disproved that result; more in detail, we showed that if we consider only the top ranked systems according to MAP (i.e., the systems for which evaluation is more interesting) there is no evidence that the ranking is affected only by easy topics; finally (iii) we proved that the above results are robust to the change of the metric used, even when GMAP, a metric that is more sensitive to low AP values, is used.

We leave plenty of space for future work. It would be interesting to investigate the effect of other metrics, like yaAP [6] which considers the number of relevant documents that can be related to topic difficulty; or NAP [3], which explicitly takes topic difficulty into consideration. It would be interesting to repeat the experiments considering a "dual" scenario, thus considering the topics and investigating the correlation between the systems and their ability to distinguish between easy and hard topics (according to some metric). This experiment might provide an explanation for the high AAP-hubness (and GAAP-hubness) correlation values in TB06. It is possible to think of a two-step evaluation of the systems (as it has already been suggested [3]): first we can use all the topics, or even the easy ones only, to evaluate the systems and provide a first rank; later, we can select the most effective systems and the most difficult topics, and use them to fine-tune the ranking of the most effective systems. This evaluation should not be affected from the undesired features that manifest with the classical one-step evaluation and has the potential of being more economic. The work on using fewer topics [1] is also to be taken into account when considering such an alternative two-step evaluation process. Finally, and more in detail, it would be interesting to investigate further about the features which make the correlation between topic ease and system effectiveness suddenly increase when considering about the 75% of the systems (sorted from more effective to less effective).

References

1. Berto, A., Mizzaro, S., Robertson, S.: On using fewer topics in information retrieval evaluations. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval, ICTIR 2013 (2013)
2. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999). http://dl.acm.org/citation.cfm?id=324140
3. Mizzaro, S.: The good, the bad, the difficult, and the easy: something wrong with information retrieval evaluation? In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 642–646. Springer, Heidelberg (2008). doi:10.1007/978-3-540-78646-7_71
4. Mizzaro, S., Robertson, S.: HITS hits TREC: exploring IR evaluation results with network analysis. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2007). http://dl.acm.org/citation.cfm?id=1277824&CFID=912758227&CFTOKEN=40185811
5. Robertson, S.: On GMAP: and other transformations. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006 (2006)
6. Robertson, S.: On smoothing average precision. In: Baeza-Yates, R., Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 158–169. Springer, Heidelberg (2012). doi:10.1007/978-3-642-28997-2_14