HITS Hits TREC — Exploring IR Evaluation Results with Network Analysis

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine,
Udine, Italy
mizzaro@dimi.uniud.it

Stephen Robertson
Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK
ser@microsoft.com

ABSTRACT

We propose a novel method of analysing data gathered from TREC or similar information retrieval evaluation experiments. We define two normalized versions of average precision, that we use to construct a weighted bipartite graph of TREC systems and topics. We analyze the meaning of well known — and somewhat generalized — indicators from social network analysis on the Systems-Topics graph. We apply this method to an analysis of TREC 8 data; among the results, we find that authority measures systems performance, that hubness of topics reveals that some topics are better than others at distinguishing more or less effective systems, that with current measures a system that wants to be effective in TREC needs to be effective on easy topics, and that by using different effectiveness measures this is no longer the case.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Experimentation

Keywords

IR evaluation, TREC, Social Network Analysis, Kleinberg's HITS algorithm.

1. INTRODUCTION

Evaluation is a primary concern in the Information Retrieval (IR) field. TREC (Text REtrieval Conference) [12, 15] is an annual benchmarking exercise that has become a de facto standard in IR evaluation: before the actual conference, TREC provides to participants a collection of documents and a set of *topics* (representations of information needs). Participants use their systems to retrieve, and submit to TREC, a list of documents for each topic. After the lists have been submitted and pooled, the TREC organizers employ human assessors to provide relevance judgements on the pooled set. This defines a set of relevant documents for each topic. System effectiveness is then measured by well established metrics (Mean Average Precision being the most used). Other conferences such as NTCIR, INEX, CLEF provide comparable data.

Network analysis is a discipline that studies features and properties of (usually large) networks, or graphs. Of particular importance is Social Network Analysis [16], that studies networks made up by links among humans (friendship, acquaintance, co-authorship, bibliographic citation, etc.).

Network analysis and IR fruitfully meet in Web Search Engine implementation, as is already described in textbooks [3,6]. Current search engines use link analysis techniques to help rank the retrieved documents. Some indicators (and the corresponding algorithms that compute them) have been found useful in this respect, and are nowadays well known: Inlinks (the number of links to a Web page), PageRank [7], and HITS (Hyperlink-Induced Topic Search) [5]. Several extensions to these algorithms have been and are being proposed. These indicators and algorithms might be quite general in nature, and can be used for applications which are very different from search result ranking. One example is using HITS for stemming, as described by Agosti et al. [1].

In this paper, we propose and demonstrate a method for constructing a network, specifically a weighted complete bidirectional directed bipartite graph, on a set of TREC topics and participating systems. Links represent effectiveness measurements on system-topic pairs. We then apply analysis methods originally developed for search applications to the resulting network. This reveals phenomena previously hidden in TREC data. In passing, we also provide a small generalization to Kleinberg's HITS algorithm, as well as to Inlinks and PageRank.

The paper is organized as follows: Sect. 2 gives some motivations for the work. Sect. 3 presents the basic ideas of normalizing average precision and of constructing a systems-topics graph, whose properties are analyzed in Sect. 4; Sect. 5 presents some experiments on TREC 8 data; Sect. 6 discusses some issues and Sect. 7 closes the paper.

2. MOTIVATIONS

We are interested in the following hypotheses:

1. Some systems are more effective than others;

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

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

(a)

	t_1	t_2	$\cdots$	MAP
s_1	0.5	0.4	$\cdots$	0.6
s_2	0.4	$\cdots$	$\cdots$	0.3
$\vdots$	$\vdots$		$\ddots$	$\vdots$
AAP	0.6	0.3	$\cdots$	

(b)

Table 1: AP, MAP and AAP

2. Some topics are easier than others;
3. Some systems are better than others at distinguishing easy and difficult topics;
4. Some topics are better than others at distinguishing more or less effective systems.

The first of these hypotheses needs no further justification – every reported significant difference between any two systems supports it. There is now also quite a lot of evidence for the second, centered on the TREC Robust Track [14]. Our primary interest is in the third and fourth. The third might be regarded as being of purely academic interest; however, the fourth has the potential for being of major practical importance in evaluation studies. If we could identify a relatively small number of topics which were really good at distinguishing effective and ineffective systems, we could save considerable effort in evaluating systems.

One possible direction from this point would be to attempt direct identification of such small sets of topics. However, in the present paper, we seek instead to explore the relationships suggested by the hypotheses, between what different topics tell us about systems and what different systems tell us about topics. We seek methods of building and analysing a matrix of system-topic normalised performances, with a view to giving insight into the issue and confirming or refuting the third and fourth hypotheses. It turns out that the obvious symmetry implied by the above formulation of the hypotheses is a property worth investigating, and the investigation does indeed give us valuable insights.

3. THE IDEA

3.1 1st step: average precision table

From TREC results, one can produce an Average Precision (AP) table (see Tab. 1a): each $\text{AP}(s_i, t_j)$ value measures the AP of system s_i on topic t_j.

Besides AP values, the table shows *Mean Average Precision* (MAP) values i.e., the mean of the AP values for a single system over all topics, and what we call *Average Average Precision* (AAP) values i.e., the average of the AP values for a single topic over all systems:

$$\text{MAP}(s_i) = \frac{1}{n}\sum_{j=1}^{n}\text{AP}(s_i, t_j), \tag{1}$$

$$\text{AAP}(t_j) = \frac{1}{m}\sum_{i=1}^{m}\text{AP}(s_i, t_j). \tag{2}$$

MAPs are indicators of systems performance: higher MAP means *good* system. AAP are indicators of the performance on a topic: higher AAP means *easy* topic — a topic on which all or most systems have good performance.

3.2 Critique of pure AP

MAP is a standard, well known, and widely used IR effectiveness measure. Single AP values are used too (e.g., in AP histograms). Topic difficulty is often discussed (e.g., in TREC Robust track [14]), although AAP values are not used and, to the best of our knowledge, have never been proposed (the *median*, not the average, of AP on a topic is used to produce TREC AP histograms [11]). However, the AP values in Tab. 1 present two limitations, which are symmetric in some respect:

- **Problem 1**. They are not reliable to compare the effectiveness of a system on different topics, *relative to the other systems*. If, for example, $\text{AP}(s_1, t_1) > \text{AP}(s_1, t_2)$, can we infer that s_1 is a good system (i.e., has a good performance) on t_1 and a bad system on t_2? The answer is no: t_1 might be an easy topic (with high AAP) and t_2 a difficult one (low AAP). See an example in Tab. 1b: s_1 is outperformed (on average) by the other systems on t_1, and it outperforms the other systems on t_2.
- **Problem 2**. Conversely, if, for example, $\text{AP}(s_1, t_1) > \text{AP}(s_2, t_1)$, can we infer that t_1 is considered easier by s_1 than by s_2? No, we cannot: s_1 might be a good system (with high MAP) and s_2 a bad one (low MAP); see an example in Tab. 1b.

These two problems are a sort of breakdown of the well known high influence of topics on IR evaluation; again, our formulation makes explicit the topics / systems symmetry.

3.3 2nd step: normalizations

To avoid these two problems, we can normalize the AP table in two ways. The first normalization removes the influence of the single topic ease on system performance. Each $\text{AP}(s_i, t_j)$ value in the table depends on both system goodness and topic ease (the value will increase if a system is good and/or the topic is easy). By subtracting from each $\text{AP}(s_i, t_j)$ the $\text{AAP}(t_j)$ value, we obtain "normalized" AP values ($\overline{\text{AP}_\text{A}}(s_i, t_j)$, *Normalized AP according to AAP*):

$$\overline{\text{AP}_\text{A}}(s_i, t_j) = \text{AP}(s_i, t_j) - \text{AAP}(t_j), \tag{3}$$

that depend on system performance only (the value will increase only if system performance is good). See Tab. 2a.

The second normalization removes the influence of the single system effectiveness on topic ease: by subtracting from each $\text{AP}(s_i, t_j)$ the $\text{MAP}(s_i)$ value, we obtain "normalized" AP values ($\overline{\text{AP}_\text{M}}(s_i, t_j)$, *Normalized AP according to MAP*):

$$\overline{\text{AP}_\text{M}}(s_i, t_j) = \text{AP}(s_i, t_j) - \text{MAP}(s_i), \tag{4}$$

that depend on topic ease only (the value will increase only if the topic is easy, i.e., all systems perform well on that topic). See Tab. 2b.

In other words, $\overline{\text{AP}_\text{A}}$ avoids Problem 1: $\overline{\text{AP}_\text{A}}(s, t)$ values measure the performance of system s on topic t normalized

	t_1	$\cdots$	t_n	$\overline{\text{MAP}}$
s_1	$\overline{\text{AP}_\text{A}}(s_1,t_1)$	$\cdots$	$\overline{\text{AP}_\text{A}}(s_1,t_n)$	$\overline{\text{MAP}}(s_1)$
$\vdots$		$\ddots$		$\vdots$
s_m	$\overline{\text{AP}_\text{A}}(s_m,t_1)$	$\cdots$	$\overline{\text{AP}_\text{A}}(s_m,t_n)$	$\overline{\text{MAP}}(s_m)$
	0	$\cdots$	0	0

(a)

	t_1	$\cdots$	t_n	
s_1	$\overline{\text{AP}_\text{M}}(s_1,t_1)$	$\cdots$	$\overline{\text{AP}_\text{M}}(s_1,t_n)$	0
$\vdots$		$\ddots$		$\vdots$
s_m	$\overline{\text{AP}_\text{M}}(s_m,t_1)$	$\cdots$	$\overline{\text{AP}_\text{M}}(s_m,t_n)$	0
$\overline{\text{AAP}}$	$\overline{\text{AAP}}(t_1)$	$\cdots$	$\overline{\text{AAP}}(t_n)$	0

(b)

	t_1	t_2	$\cdots$	$\overline{\text{MAP}}$
s_1	-0.1	0.1	$\cdots$	$\ldots$
s_2	0.2	$\cdots$	$\cdots$	$\ldots$
$\vdots$			$\ddots$	$\vdots$
	0	0	$\cdots$	

(c)

	t_1	t_2	$\cdots$	
s_1	-0.1	-0.2	$\cdots$	0
s_2	0.1	$\cdots$	$\cdots$	0
$\vdots$			$\ddots$	$\vdots$
$\overline{\text{AAP}}$	$\ldots$	$\ldots$	$\cdots$	

(d)

Table 2: Normalizations: $\overline{\text{AP}_\text{A}}$ and $\overline{\text{MAP}}$: normalized AP ($\overline{\text{AP}_\text{A}}$) and MAP ($\overline{\text{MAP}}$) (a); normalized AP ($\overline{\text{AP}_\text{M}}$) and AAP ($\overline{\text{AAP}}$) (b); a numeric example (c) and (d)

according to the ease of the topic (easy topics will *not* have higher $\overline{\text{AP}_\text{A}}$ values). Now, if, for example, $\overline{\text{AP}_\text{A}}(s_1,t_2) > \overline{\text{AP}_\text{A}}(s_1,t_1)$, we can infer that s_1 is a good system on t_2 and a bad system on t_1 (see Tab. 2c). Vice versa, $\overline{\text{AP}_\text{M}}$ avoids Problem 2: $\overline{\text{AP}_\text{M}}(s,t)$ values measure the ease of topic t according to system s, normalized according to goodness of the system (good systems will *not* lead to higher $\overline{\text{AP}_\text{M}}$ values). If, for example, $\overline{\text{AP}_\text{M}}(s_2,t_1) > \overline{\text{AP}_\text{M}}(s_1,t_1)$, we can infer that t_1 is considered easier by s_2 than by s_1 (see Tab. 2d).

On the basis of Tables 2a and 2b, we can also define two new measures of system effectiveness and topic ease, i.e., a *Normalized MAP* ($\overline{\text{MAP}}$), obtained by averaging the $\overline{\text{AP}_\text{A}}$ values on one row in Tab. 2a, and a *Normalized AAP* ($\overline{\text{AAP}}$), obtained by averaging the $\overline{\text{AP}_\text{M}}$ values on one column in Tab. 2b:

$$\overline{\text{MAP}}(s_i) = \frac{1}{n}\sum_{j=1}^{n}\overline{\text{AP}_\text{A}}(s_i,t_j) \tag{5}$$

$$\overline{\text{AAP}}(t_j) = \frac{1}{m}\sum_{i=1}^{m}\overline{\text{AP}_\text{M}}(s_i,t_j). \tag{6}$$

Thus, overall system performance can be measured, besides by means of MAP, also by means of $\overline{\text{MAP}}$. Moreover, $\overline{\text{MAP}}$ is equivalent to MAP, as can be immediately proved by using Eqs. (5), (3), and (1):

$$\overline{\text{MAP}}(s_i) = \frac{1}{n}\sum_{j=1}^{n}(\text{AP}(s_i,t_j) - \text{AAP}(t_j)) =$$

$$= \text{MAP}(s_i) - \frac{1}{n}\sum_{j=1}^{n}\text{AAP}(t_j)$$

(and $\frac{1}{n}\sum_{j=1}^{n}\text{AAP}(t_j)$ is the same for all systems). And, conversely, overall topic ease can be measured, besides by means of AAP, also by means of $\overline{\text{AAP}}$, and this is equivalent (the proof is analogous, and relies on Eqs. (6), (4), and (2)).

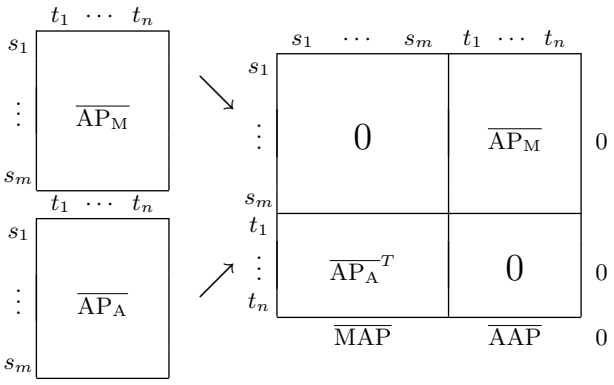

Figure 1: Construction of the adjacency matrix. $\overline{\text{AP}_\text{A}}^T$ is the transpose of $\overline{\text{AP}_\text{A}}$.

The two Tables 2a and 2b are interesting per se, and can be analyzed in several different ways. In the following we propose an analysis based on network analysis techniques, mainly Kleinberg's HITS algorithm. There is a little further discussion of these normalizations in Sect. 6.

3.4 3rd step: Systems-Topics Graph

The two tables 2a and 2b can be merged into a single one with the procedure shown in Fig. 1. The obtained matrix can be interpreted as the adjacency matrix of a complete weighted bipartite graph, that we call *Systems-Topics graph*. Arcs and weights in the graph can be interpreted as follows:

- (weight on) arc $s \rightarrow t$: how much the system s "thinks" that the topic t is easy — assuming that a system has no knowledge of the other systems (or in other words, how easy *we* might think the topic is, knowing only the results for this one system). This corresponds to $\overline{\text{AP}_\text{M}}$ values, i.e., to normalized topic ease (Fig. 2a).
- (weight on) arc $s \leftarrow t$: how much the topic t "thinks" that the system s is good — assuming that a topic has no knowledge of the other topics (or in other words, how good *we* might think the system is, knowing only the results for this one topic). This corresponds to $\overline{\text{AP}_\text{A}}$ (normalized system effectiveness, Fig. 2b).

Figs. 2c and 2d show the Systems-Topics complete weighted bipartite graph, on a toy example with 4 systems and 2 topics; the graph is split in two parts to have an understandable graphical representation: arcs in Fig. 2c are labeled with $\overline{\text{AP}_\text{M}}$ values; arcs in Fig. 2d are labeled with $\overline{\text{AP}_\text{A}}$ values.

4. ANALYSIS OF THE GRAPH

4.1 Weighted Inlinks, Outlinks, PageRank

The sum of weighted outlinks, i.e., the sum of the weights on the outgoing arcs from each node, is always zero:

- The outlinks on each node corresponding to a system s (Fig. 2c) is the sum of all the corresponding $\overline{\text{AP}_\text{M}}$ values on one row of the matrix in Tab. 2b.
- The outlinks on each node corresponding to a topic t (Fig. 2d) is the sum of all the corresponding $\overline{\text{AP}_\text{A}}$

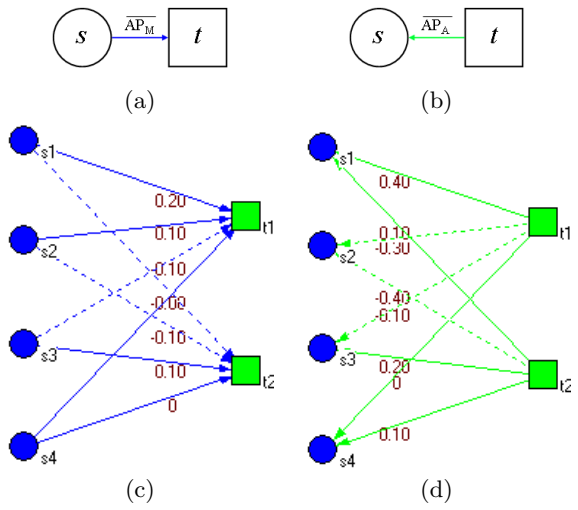

Figure 2: The relationships between systems and topics (a) and (b); and the Systems-Topics graph for a toy example (c) and (d). Dashed arcs correspond to negative values.

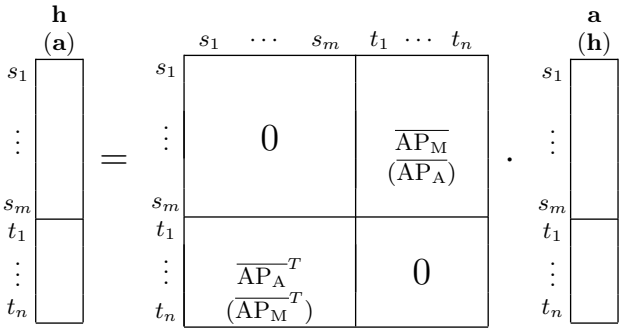

Figure 3: Hub and Authority computation

values on one row of the transpose of the matrix in Tab. 2a.

The average[1] of weighted inlinks is:

- $\overline{MAP}$ for each node corresponding to a system s; this corresponds to the average of all the corresponding $\overline{AP_A}$ values on one column of the $\overline{AP_A}^T$ part of the adjacency matrix (see Fig. 1).
- $\overline{AAP}$ for each node corresponding to a topic t; this corresponds to the average of all the corresponding $\overline{AP_M}$ values on one column of the $\overline{AP_M}$ part of the adjacency matrix (see Fig. 1).

Therefore, weighted inlinks measure either system effectiveness or topic ease; weighted outlinks are not meaningful. We could also apply the PageRank algorithm to the network; the meaning of the PageRank of a node is not quite so obvious as Inlinks and Outlinks, but it also seems a sensible measure of either system effectiveness or topic ease: if a system is effective, it will have several incoming links with high weights ($\overline{AP_A}$); if a topic is easy it will have high weights ($\overline{AP_M}$) on the incoming links too. We will see experimental confirmation in the following.

[1]Usually, the sum of the weights on the incoming arcs to each node is used in place of the average; since the graph is complete, it makes no difference.

4.2 Hubs and Authorities

Let us now turn to more sophisticated indicators. Kleinberg's HITS algorithm defines, for a directed graph, two indicators: *hubness* and *authority*; we reiterate here some of the basic details of the HITS algorithm in order to emphasize both the nature of our generalization and the interpretation of the HITS concepts in this context. Usually, hubness and authority are defined as $h(x) = \sum_{x \to y} a(y)$ and $a(x) = \sum_{y \to x} h(y)$, and described intuitively as "a good hub links many good authorities; a good authority is linked from many good hubs". As it is well known, an equivalent formulation in linear algebra terms is (see also Fig. 3):

$$\mathbf{h} = \mathbf{A}\mathbf{a} \text{ and } \mathbf{a} = \mathbf{A}^T\mathbf{h} \quad (7)$$

(where $\mathbf{h}$ is the hubness vector, with the hub values for all the nodes; $\mathbf{a}$ is the authority vector; $\mathbf{A}$ is the adjacency matrix of the graph; and $\mathbf{A}^T$ its transpose). Usually, $\mathbf{A}$ contains 0s and 1s only, corresponding to presence and absence of unweighted directed arcs, but Eq. (7) can be immediately generalized to (in fact, it is already valid for) $\mathbf{A}$ containing any real value, i.e., to weighted graphs.

Therefore we can have a "generalized version" (or rather a generalized interpretation, since the formulation is still the original one) of hubness and authority for all nodes in a graph. An intuitive formulation of this *generalized HITS* is still available, although slightly more complex: "a good hub links, by means of arcs having high weights, many good authorities; a good authority is linked, by means of arcs having high weights, from many good hubs". Since arc weights can be, in general, negative, hub and authority values can be negative, and one could speak of *unhubness* and *unauthority*; the intuitive formulation could be completed by adding that "a good hub links good unauthorities by means of links with highly negative weights; a good authority is linked by good unhubs by means of links with highly negative weights". And, also, "a good unhub links positively good unauthorities and negatively good authorities; a good unauthority is linked positively from good unhubs and negatively from good hubs".

Let us now apply generalized HITS to our Systems-Topics graph. We compute $\mathbf{a}(s)$, $\mathbf{h}(s)$, $\mathbf{a}(t)$, and $\mathbf{h}(t)$. Intuitively, we expect that $\mathbf{a}(s)$ is somehow similar to Inlinks, so it should be a measure of either systems effectiveness or topic ease. Similarly, hubness should be more similar to Outlinks, thus less meaningful, although the interplay between hub and authority might lead to the discovery of something different. Let us start by remarking that authority of topics and hubness of systems depend only on each other; similarly hubness of topics and authority of systems depend only on each other: see Figs. 2c, 2d and 3.

Thus the two graphs in Figs. 2c and 2d can be analyzed independently. In fact the entire HITS analysis could be done in one direction only, with just $\overline{AP_M}(s,t)$ values or alternatively with just $\overline{AP_A}(s,t)$. As discussed below, probably most interest resides in the hubness of topics and the authority of systems, so the latter makes sense. However, in this paper, we pursue both analyses together, because the symmetry itself is interesting.

Considering Fig. 2c we can state that:

- Authority $\mathbf{a}(t)$ of a topic node t increases when:
  - if $\mathbf{h}(s_i) > 0$, $\overline{\mathrm{AP_M}}(s_i, t)$ increases (or if $\overline{\mathrm{AP_M}}(s_i, t) > 0$, $\mathbf{h}(s_i)$ increases);
  - if $\mathbf{h}(s_i) < 0$, $\overline{\mathrm{AP_M}}(s_i, t)$ decreases (or if $\overline{\mathrm{AP_M}}(s_i, t) < 0$, $\mathbf{h}(s_i)$ decreases).
- Hubness $\mathbf{h}(s)$ of a system node s increases when:
  - if $\mathbf{a}(t_j) > 0$, $\overline{\mathrm{AP_M}}(s, t_j)$ increases (or, if $\overline{\mathrm{AP_M}}(s, t_j) > 0$, $\mathbf{a}(t_j)$ increases);
  - if $\mathbf{a}(t_j) < 0$, $\overline{\mathrm{AP_M}}(s, t_j)$ decreases (or, if $\overline{\mathrm{AP_M}}(s, t_j) < 0$, $\mathbf{a}(t_j)$ decreases).

We can summarize this as: $\mathbf{a}(t)$ is high if $\overline{\mathrm{AP_M}}(s, t)$ is high for those systems with high $\mathbf{h}(s)$; $\mathbf{h}(s)$ is high if $\overline{\mathrm{AP_M}}(s, t)$ is high for those topics with high $\mathbf{a}(t)$. Intuitively, authority $\mathbf{a}(t)$ of a topic measures topic ease; hubness $\mathbf{h}(s)$ of a system measures system's "capability" to recognize easy topics. A system with high unhubness (negative hubness) would tend to regard easy topics as hard and hard ones as easy.

The situation for Fig. 2d, i.e., for $\mathbf{a}(s)$ and $\mathbf{h}(t)$, is analogous. Authority $\mathbf{a}(s)$ of a system node s measures system effectiveness: it increases with the weight on the arc (i.e., $\overline{\mathrm{AP_A}}(s, t_j)$) and the hubness of the incoming topic nodes t_j. Hubness $\mathbf{h}(t)$ of a topic node t measures topic capability to recognize effective systems: if $\mathbf{h}(t) > 0$, it increases further if $\overline{\mathrm{AP_A}}(s, t_j)$ increases; if $\mathbf{h}(t) < 0$, it increases if $\overline{\mathrm{AP_A}}(s, t_j)$ decreases.

Intuitively, we can state that "A system has a higher authority if it is more effective on topics with high hubness"; and "A topic has a higher hubness if it is easier for those systems which are more effective in general". Conversely for system hubness and topic authority: "A topic has a higher authority if it is easier on systems with high hubness"; and "A system has a higher hubness if it is more effective for those topics which are easier in general".

Therefore, for each system we have two indicators: authority ($\mathbf{a}(s)$), measuring system effectiveness, and hubness ($\mathbf{h}(s)$), measuring system capability to estimate topic ease. And for each topic, we have two indicators: authority ($\mathbf{a}(t)$), measuring topic ease, and hubness ($\mathbf{h}(t)$), measuring topic capability to estimate systems effectiveness. We can define them formally as

$$\mathbf{a}(s) = \sum_t \mathbf{h}(t) \cdot \overline{\mathrm{AP_A}}(s, t), \quad \mathbf{h}(t) = \sum_s \mathbf{a}(s) \cdot \overline{\mathrm{AP_A}}(s, t),$$

$$\mathbf{a}(t) = \sum_s \mathbf{h}(s) \cdot \overline{\mathrm{AP_M}}(s, t), \quad \mathbf{h}(s) = \sum_t \mathbf{a}(t) \cdot \overline{\mathrm{AP_M}}(s, t).$$

We observe that the hubness of topics may be of particular interest for evaluation studies. It may be that we can evaluate the effectiveness of systems efficiently by using relatively few high-hubness topics.

5. EXPERIMENTS

We now turn to discuss if these indicators are meaningful and useful in practice, and how they correlate with standard measures used in TREC. We have built the Systems-Topics graph for TREC 8 data (featuring 128[2] systems — actually, runs — on 50 topics). This section illustrates the results obtained mining these data according to the method presented in previous sections.

[2]Actually, TREC 8 data features 129 systems; due to some bug in our scripts, we did not include one system (8manexT3D1N0), but the results should not be affected.

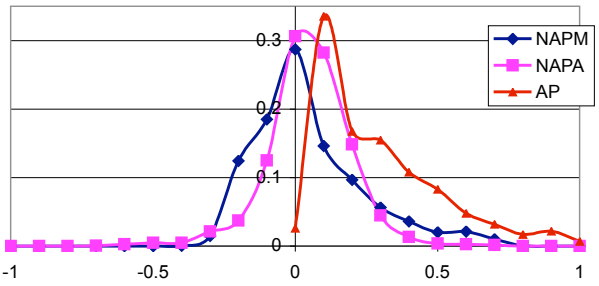

Figure 4: Distributions of AP, $\overline{\mathrm{AP_A}}$, and $\overline{\mathrm{AP_M}}$ values in TREC 8 data

	MAP	In	PR	H	A
MAP	1	1.0	1.0	.80	.99
Inlinks		1	1.0	.80	.99
PageRank			1	.80	.99
Hub				1	.87

(a)

	AAP	In	PR	H	A
AAP	1	1.0	1.0	.92	1.0
Inlinks		1	1.0	.92	1.0
PageRank			1	.92	1.0
Hub				1	.93

(b)

Table 3: Correlations between network analysis measures and MAP (a) and AAP (b)

Fig. 4 shows the distributions of AP, $\overline{\mathrm{AP_A}}$, and $\overline{\mathrm{AP_M}}$: whereas AP is very skewed, both $\overline{\mathrm{AP_A}}$ and $\overline{\mathrm{AP_M}}$ are much more symmetric (as it should be, since they are constructed by subtracting the mean). Tables 3a and 3b show the Pearson's correlation values between Inlinks, PageRank, Hub, Authority and, respectively, MAP or AAP (Outlinks values are not shown since they are always zero, as seen in Sect. 4). As expected, Inlinks and PageRank have a perfect correlation with MAP and AAP. Authority has a very high correlation too with MAP and AAP; Hub assumes slightly lower values.

Let us analyze the correlations more in detail. The correlations chart in Figs. 5a and 5b demonstrate the high correlation between Authority and MAP or AAP. Hubness presents interesting phenomena: both Fig. 5c (correlation with MAP) and Fig. 5d (correlation with AAP) show that correlation is not exact, but neither is it random. This, given the meaning of hubness (capability in estimating topic ease and system effectiveness), means two things: (i) more effective systems are better at estimating topic ease; and (ii) easier topics are better at estimating system effectiveness. Whereas the first statement is fine (there is nothing against it), the second 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, would be an important result for IR effectiveness; con-

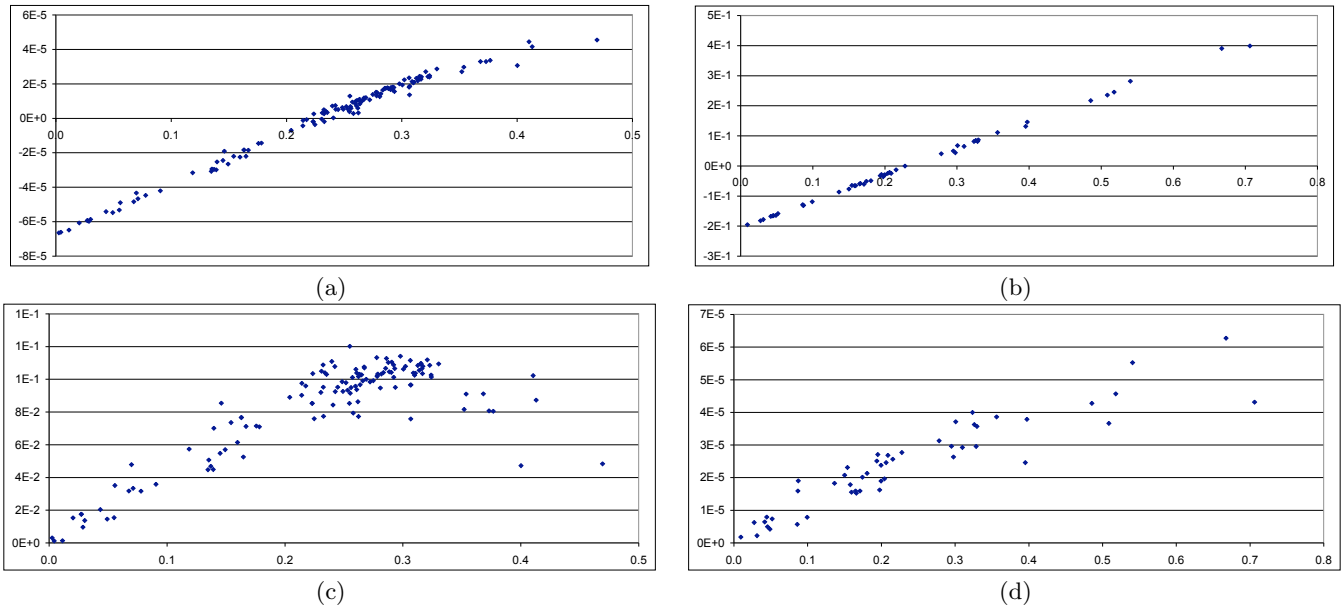

Figure 5: Correlations: MAP (x axis) and authority (y axis) of systems (a); AAP and authority of topics (b); MAP and hub of systems (c) and AAP and hub of topics (d)

versely, a system capable of performing well on easy topics is just a confirmation of the state of the art. Indeed, the correlation between hubness and AAP (statement (i) above) is higher than the correlation between hubness and MAP (corresponding to statement (ii)): 0.92 vs. 0.80. However, this phenomenon is quite strong. This is also confirmed by the work being done on the TREC Robust Track [14].

In this respect, it is interesting to see what happens if we use a different measure from MAP (and AAP). The GMAP (Geometric MAP) metric is defined as the geometric mean of AP values, or equivalently as the arithmetic mean of the logarithms of AP values [8]. GMAP has the property of giving more weight to the low end of the AP scale (i.e., to low AP values), and this seems reasonable, since, intuitively, a performance increase in MAP values from 0.01 to 0.02 should be more important than an increase from 0.81 to 0.82. To use GMAP in place of MAP and AAP, we only need to take the logarithms of initial AP values, i.e., those in Tab. 1a (zero values are modified into $\varepsilon = 0.00001$). We then repeat the same normalization process (with GMAP and GAAP — Geometric AAP — replacing MAP and AAP): whereas authority values still perfectly correlate with GMAP (0.99) and GAAP (1.00), the correlation with hubness largely disappears (values are -0.16 and -0.09 – slightly negative but not enough to concern us). This is yet another confirmation that TREC effectiveness as measured by MAP depends mainly on easy topics; GMAP appears to be a more balanced measure. Note that, perhaps surprisingly, GMAP is indeed fairly well balanced, not biased in the opposite direction – that is, it does not over-emphasize the difficult topics.

In Sect. 6.3 below, we discuss another transformation, replacing the log function used in GMAP with logit. This has a similar effect: the correlation of mean logitAP and average logitAP with hubness are now small positive numbers (0.23 and 0.15 respectively), still comfortably away from the high correlations with regular MAP and AAP, i.e., not presenting the problematic phenomenon (ii) above (over-dependency on easy topics).

We also observe that hub values are positive, whereas authority assumes, as predicted, both positive and negative values. An intuitive justification is that negative hubness would indicate a node which disagrees with the other nodes, e.g., a system which does better on difficult topics, or a topic on which bad systems do better; such systems and topics would be quite strange, and probably do not appear in TREC. Finally, although one might think that topics with several relevant documents are more important and difficult, this is not the case: there is no correlation between hub (or any other indicator) and the number of documents relevant to a topic.

6. DISCUSSION

6.1 Related work

There has been considerable interest in recent years in questions of statistical significance of effectiveness comparisons between systems (e.g. [2, 9]), and related questions of how many topics might be needed to establish differences (e.g. [13]). We regard some results of the present study as in some way complementary to this work, in that we make a step towards answering the question "*Which* topics are best for establishing differences?".

The results on evaluation without relevance judgements such as [10] show that, to some extent, good systems agree on which are the good documents. We have not addressed the question of individual documents in the present analysis, but this effect is certainly analogous to our results.

6.2 Are normalizations necessary?

At this point it is also worthwhile to analyze what would happen without the MAP- and AAP-normalizations defined in Sect. 3.3. Indeed, the process of graph construction (Sect. 3.4) is still valid: both the $\overline{\mathrm{AP}_\mathrm{M}}$ and $\overline{\mathrm{AP}_\mathrm{A}}$ matrices are replaced by the AP one, and then everything goes on as

above. Therefore, one might think that the normalizations are unuseful in this setting.

This is not the case. From the theoretical point of view, the AP-only graph does not present the interesting properties above discussed: since the AP-only graph is symmetrical (the weight on each incoming link is equal to the weight on the corresponding outgoing link), Inlinks and Outlinks assume the same values. There is symmetry also in computing hub and authority, that assume the same value for each node since the weights on the incoming and outgoing arcs are the same. This could be stated in more precise and formal terms, but one might still wonder if on the overall graph there are some sort of counterbalancing effects. It is therefore easier to look at experimental data, which confirm that the normalizations are needed: the correlations between AP, Inlinks, Outlinks, Hub, and/or Authority are *all* very close to one (none of them is below 0.98).

6.3 Are these normalizations sufficient?

It might be argued that (in the case of $\overline{\mathrm{AP}_\mathrm{A}}$, for example) the amount we have subtracted from each AP value is topic-dependent, therefore the range of the resulting $\overline{\mathrm{AP}_\mathrm{A}}$ value is also topic-dependent (e.g. the maximum is $1 - \mathrm{AAP}(t_j)$ and the minimum is $-\mathrm{AAP}(t_j)$). This suggests that the cross-topic comparisons of these values suggested in Sect. 3.3 may not be reliable. A similar issue arises for $\overline{\mathrm{AP}_\mathrm{M}}$ and comparisons across systems.

One possible way to overcome this would be to use an unconstrained measure whose range is the full real line. Note that in applying the method to GMAP by using $\log \mathrm{AP}$, we avoid the problem with the lower limit but retain it for the upper limit. One way to achieve an unconstrainted range would be to use the logit function rather than the log [4, 8].

We have also run this variant (as already reported in Sect. 5 above), and it appears to provide very similar results to the GMAP results already given. This is not surprising, since in practice the two functions are very similar over most of the operative range. The normalizations thus seem reliable.

6.4 On $\mathbf{A}\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{T}\mathbf{A}$

It is well known that $\mathbf{h}$ and $\mathbf{a}$ vectors are the principal left eigenvectors of $\mathbf{A}\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{T}\mathbf{A}$, respectively (this can be easily derived from Eqs. (7)), and that, in the case of citation graphs, $\mathbf{A}\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{T}\mathbf{A}$ represent, respectively, bibliographic coupling and co-citations. What is the meaning, if any, of $\mathbf{A}\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{T}\mathbf{A}$ in our Systems-Topics graph? It is easy to derive that:

$$\mathbf{A}\mathbf{A}^\mathrm{T}[i,j] = \begin{cases} 0 & \begin{cases}\text{if } i \in S \wedge j \in T \\ \text{or } i \in T \wedge j \in S\end{cases} \\ \sum_k \mathbf{A}[i,k] \cdot \mathbf{A}[j,k] & \text{otherwise} \end{cases}$$

$$\mathbf{A}^\mathrm{T}\mathbf{A}[i,j] = \begin{cases} 0 & \begin{cases}\text{if } i \in S \wedge j \in T \\ \text{or } i \in T \wedge j \in S\end{cases} \\ \sum_k \mathbf{A}[k,i] \cdot \mathbf{A}[k,j] & \text{otherwise} \end{cases}$$

(where S is the set of indices corresponding to systems and T the set of indices corresponding to topics). Thus $\mathbf{A}\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{T}\mathbf{A}$ are block diagonal matrices, with two blocks each, one relative to systems and one relative to topics:

(a) if $i, j \in S$, then $\mathbf{A}\mathbf{A}^\mathrm{T}[i,j] = \sum_{k \in T} \overline{\mathrm{AP}_\mathrm{M}}(i,k) \cdot \overline{\mathrm{AP}_\mathrm{M}}(j,k)$ measures how much the two systems i and j agree in estimating topics ease ($\overline{\mathrm{AP}_\mathrm{M}}$): high values mean that the two systems agree on topics ease.

(b) if $i, j \in T$, then $\mathbf{A}\mathbf{A}^\mathrm{T}[i,j] = \sum_{k \in S} \overline{\mathrm{AP}_\mathrm{A}}(k,i) \cdot \overline{\mathrm{AP}_\mathrm{A}}(k,j)$ measures how much the two topics i and j agree in estimating systems effectiveness ($\overline{\mathrm{AP}_\mathrm{A}}$): high values mean that the two topics agree on systems effectiveness (and that TREC results would not change by leaving out one of the two topics).

(c) if $i, j \in S$, then $\mathbf{A}^\mathrm{T}\mathbf{A}[i,j] = \sum_{k \in T} \overline{\mathrm{AP}_\mathrm{A}}(i,k) \cdot \overline{\mathrm{AP}_\mathrm{A}}(j,k)$ measures how much agreement on the effectiveness of two systems i and j there is over all topics: high values mean that many topics quite agree on the two systems effectiveness; low values single out systems that are somehow controversial, and that need several topics to have a correct effectiveness assessment.

(d) if $i, j \in T$, then $\mathbf{A}^\mathrm{T}\mathbf{A}[i,j] = \sum_{k \in S} \overline{\mathrm{AP}_\mathrm{M}}(k,i) \cdot \overline{\mathrm{AP}_\mathrm{M}}(k,j)$ measures how much agreement on the ease of the two topics i and j there is over all systems: high values mean that many systems quite agree on the two topics ease.

Therefore, these matrices are meaningful and somehow interesting. For instance, the submatrix (b) corresponds to a weighted undirected complete graph, whose nodes are the topics and whose arc weights are a measure of how much two topics agree on systems effectiveness. Two topics that are very close on this graph give the same information, and therefore one of them could be discarded without changes in TREC results. It would be interesting to cluster the topics on this graph. Furthermore, the matrix/graph (a) could be useful in TREC pool formation: systems that do not agree on topic ease would probably find different relevant documents, and should therefore be complementary in pool formation. Note that no notion of single documents is involved in the above analysis.

6.5 Insights

As indicated, the primary contribution of this paper has been a method of analysis. However, in the course of applying this method to one set of TREC results, we have achieved some insights relating to the hypotheses formulated in Sect. 2:

- We confirm Hypothesis 2 above, that some topics are easier than others.
- Differences in the hubness of systems reveal that some systems are better than others at distinguishing easy and difficult topics; thus we have some confirmation of Hypothesis 3.
- There are some relatively idiosyncratic systems which do badly on some topics generally considered easy but well on some hard topics. However, on the whole, the more effective systems are better at distinguishing easy and difficult topics. This is to be expected: a really bad system will do badly on everything, while even a good system may have difficulty with some topics.
- Differences in the hubness of topics reveal that some topics are better than others at distinguising more or less effective systems; thus we have some confirmation of Hypothesis 4.

- If we use MAP as the measure of effectiveness, it is also true that the easiest topics are better at distinguishing more or less effective systems. As argued in Sect. 5, this is an undesirable property. GMAP is more balanced.

Clearly these ideas need to be tested on other data sets. However, they reveal that the method of analysis proposed in this paper can provide valuable information.

6.6 Selecting topics

The confirmation of Hypothesis 4 leads, as indicated, to the idea that we could do reliable system evaluation on a much smaller set of topics, provided we could select such an appropriate set. This selection may not be straightforward, however. It is possible that simply selecting the high hubness topics will achieve this end; however, it is also possible that there are significant interactions between topics which would render such a simple rule ineffective. This investigation would therefore require serious experimentation. For this reason we have not attempted in this paper to point to the specific high hubness topics as being good for evaluation. This is left for future work.

7. CONCLUSIONS AND FUTURE DEVELOPMENTS

The contribution of this paper is threefold:

- we propose a novel way of normalizing AP values;
- we propose a novel method to analyse TREC data;
- the method applied on TREC data does indeed reveal some hidden properties.

More particularly, we propose Average Average Precision (AAP), a measure of topic ease, and a novel way of normalizing the average precision measure in TREC, on the basis of both MAP (Mean Average Precision) and AAP. The normalized measures ($\overline{\text{AP}_\text{M}}$ and $\overline{\text{AP}_\text{A}}$) are used to build a bipartite weighted Systems-Topics graph, that is then analyzed by means of network analysis indicators widely known in the (social) network analysis field, but somewhat generalised. We note that no such approach to TREC data analysis has been proposed so far. The analysis shows that, with current measures, a system that wants to be effective in TREC needs to be effective on easy topics. Also, it is suggested that a cluster analysis on topic similarity can lead to relying on a lower number of topics.

Our method of analysis, as described in this paper, can be applied only *a posteriori*, i.e., once we have all the topics and all the systems available. Adding (removing) a new system / topic would mean re-computing hubness and authority indicators. Moreover, we are not explicitly proposing a change to current TREC methodology, although this could be a by-product of these — and further — analyses.

This is an initial work, and further analyses could be performed. For instance, other effectiveness metrics could be used, in place of AP. Other centrality indicators, widely used in social network analysis, could be computed, although probably with similar results to PageRank. It would be interesting to compute the higher-order eigenvectors of $\mathbf{A}^\text{T}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\text{T}$. The same kind of analysis could be performed at the document level, measuring document ease. Hopefully, further analyses of the graph defined in this paper, according to the approach described, can be insightful for a better understanding of TREC or similar data.

Acknowledgments

We would like to thank Nick Craswell for insightful discussions and the anonymous referees for useful remarks. Part of this research has been carried on while the first author was visiting Microsoft Research Cambridge, whose financial support is acknowledged.

8. REFERENCES

[1] M. Agosti, M. Bacchin, N. Ferro, and M. Melucci. Improving the automatic retrieval of text documents. In *Proceedings of the 3rd CLEF Workshop*, volume 2785 of *LNCS*, pages 279–290, 2003.

[2] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *23rd SIGIR*, pages 33–40, 2000.

[3] S. Chakrabarti. *Mining the Web*. Morgan Kaufmann, 2003.

[4] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *29th SIGIR*, pages 533–540, 2006.

[5] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. of the ACM*, 46(5):604–632, 1999.

[6] M. Levene. *An Introduction to Search Engines and Web Navigation*. Addison Wesley, 2006.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web, 1998. `http://dbpubs.stanford.edu:8090/pub/1999-66`.

[8] S. Robertson. On GMAP – and other transformations. In *13th CIKM*, pages 78–83, 2006.

[9] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *28th SIGIR*, pages 162–169, 2005. `http://doi.acm.org/10.1145/1076034.1076064`.

[10] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *24th SIGIR*, pages 66–73, 2001.

[11] TREC Common Evaluation Measures, 2005. `http://trec.nist.gov/pubs/trec14/appendices/CE.MEASURES05.pdf` (Last visit: Jan. 2007).

[12] Text REtrieval Conference (TREC). `http://trec.nist.gov/` (Last visit: Jan. 2007).

[13] E. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *25th SIGIR*, pages 316–323, 2002.

[14] E. M. Voorhees. Overview of the TREC 2005 Robust Retrieval Track. In *TREC 2005 Proceedings*, 2005.

[15] E. M. Voorhees and D. K. Harman. *TREC — Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[16] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, UK, 1994.