Knowledge and Information Systems manuscript No.
(will be inserted by the editor)

How many performance measures to evaluate Information Retrieval Systems?

Alain Baccini · Sébastien Déjean · Laetitia Lafage · Josiane Mothe

Received: date / Accepted: date

Abstract Evaluating effectiveness of information retrieval systems is achieved by performing on a collection of documents, a search, in which a set of test queries are performed and, for each query, the list of the relevant documents. This evaluation framework also includes performance measures making it possible to control the impact of a modification of search parameters. The program trec_eval calculates a large number of measures, some being more used like the mean average precision or recall-precision curves. The motivation of our work is to compare all measures and to help the user to choose a small number of them when evaluating different information retrieval systems. In this paper, we present the study we carried out from a massive data analysis of TREC results. Relationships between the 130 measures calculated by trec_eval for individual queries are investigated and we show that they can be clustered into homogeneous clusters.

Keywords information retrieval · performance measures · evaluation · statistical data analysis.

1 Introduction

Evaluating effectiveness of information retrieval systems (IRS) performances is a core problem in IR. The evaluation model commonly used today is based on the model developed in the Cranfield project [12]. This model implies a collection of documents on which a search is carried out, a set of queries and the list of the relevant documents for each query. This evaluation model also includes performance (or evaluation) measures

A. Baccini, S. Déjean
Institut de Mathématiques de Toulouse UMR 5219
Université de Toulouse et CNRS
E-mail: alain.baccini, sebastien.dejean@math.univ-toulouse.fr

L. Lafage, J. Mothe
Institut de Recherche en Informatique de Toulouse UMR 5505
Institut Universitaire de Formation des Maîtres
Université de Toulouse et CNRS
E-mail: laetitia.lafage, josiane.mothe@irit.fr

somehow associated with recall and precision. At least, this model makes it possible to control the impact of a modification of search parameters of an IRS, with a reasonable cost compared to an evaluation implying users to each stage [29,37].

Even if some evaluation frameworks are proposed for better taking into account the user and the acquisition of information [14,17], the Cranfield model remains the dominant approach when assessing IRS [6]. However, over the years, the model has been refined. The first aspect relates to the confidence in results obtained by various systems via statistical tests [15]. The second aspect, which is much more related to this paper, is the number of measures used over the studies. The trec_eval tool used for example in Text REtrieval Conference (TREC - trec.nist.gov), calculates a set of 135 measures (we will see that we use only 130 of them, see section 3.2). Some measures are more widely used for comparing systems and can be used in order to rank systems. We can quote that the Mean Average Precision (MAP) and the recall-precision curves are used for global comparisons [38]. Some methods are tuned in order to enhance a retrieval property [40], such as high precision for example [20]. In some other cases, it can be useful to compare two approaches or the impact of the variation of a parameter, on more than one performance measure; but using the entire set of existing measures may be too difficult to interpret. It is an issue to select a small subset of these measures. The work presented in this paper aims at answering this issue.

In this paper, we propose a methodology to study the correlations between IR performance measures in order to:

1. detect clusters of highly correlated measures;
2. define a subset of weakly correlated, and thus complementary performance measures by selecting one representative from each cluster. Indeed, measures from a given cluster are highly redundant. Thus, the subset of representative measures provides a global overview of the effectiveness of the system.

We present a deep analysis of these correlations in the case of TREC adhoc data (`trec.nist.gov`).

In order to clarify the following of the paper, we fix the idea behind the vocabulary we use. When two measures are statistically highly correlated, we claim that they are redundant: when one is given, the other one is not needed. This does not mean that they are similar. Let's take a basic example to illustrate this point. When studying morphological parameters of humans, one can measure the size of individuals as well as their leg length. It can be assumed that size and leg length are highly correlated, so if we have one measure, the second one is not needed to decide whether an individual is rather small or tall. But, of course, size and leg length are not similar; they do not measure the same thing.

The paper is organized as follows. Section 2 presents a review of the literature. Then we present the data used in this work in section 3. In order to study which measures it would be judicious to choose when comparing systems, we propose a methodology based on statistical methods presented in section 4. The results of the analysis based on 7 collections of TREC adhoc are presented in section 5. In section 6, we discuss the results and section 7 is the conclusion of this paper.

2 Related works

It is already known that performance measures more or less correlate with each other; however a few studies have been carried out in order to describe precisely the patterns of the highest correlated measures.

Tague-Sutcliffe and Blustein [34] show that R-precision and average precision are strongly correlated when considering TREC3 adhoc data; this has been confirmed, for example in [36]. Aslam et al. [3] study further these two measures and shows, using TREC8 adhoc, that this strong correlation is possibly related to the fact that two measures geometrically approximate the surface which is located below the recall precision curve.

Ishioka [16] analyzes the relationships between F-measure (which measures how the precision decreases when the recall increases), break-even point (when recall equals precision) and 11-point averaged precision, considering the coefficient Φ (Phi) on a specific 2×2 contingency table: the cross-classification table relative to relevant/nonrelevant documents in rows and retrieved/not retrieved documents in columns. The author shows that F-measure and coefficient Φ have similar properties; break-even point is almost equivalent to Φ. Buckley and Voorhees [9] calculate the pairwise correlations between 7 measures using the Kendall coefficient on TREC7. They show that the correlations are greater than 0.6 and that the highest correlation is between R-Precision and MAP. Egghe [13] studies the correlation between precision, recall, and F-measure. He shows that the evolution of precision according to recall follows a concave decreasing function whereas the evolution of recall according to fallout is a concave increasing function, where fallout corresponds to the percentage of retrieved non-relevant documents (thus, fallout + precision = 1).

Sakai [30] compares measures based on Boolean document relevance and measures based on graduated relevance; he shows that recall based on a graduated relevance is strongly correlated with average precision. Melucci [24] considers the case of the correlation between the ranking lists. Finally, Webber et al. [39] shows that P@10 is redundant with other measures.

The studies of the literature are based on a limited data set and on a reduced set of measures that are chosen a priori. On the contrary, we consider a more global solution aiming at selecting a minimal set of measures allowing one to compare two IRS; the selection of these measures is achieved from the results of the statistical study of measure redundancies.

3 Data

3.1 Runs from TREC

The evaluation program TREC (`trec.nist.gov`) started in 1992. It includes a certain number of tasks that change over years, among which the adhoc task. This task corresponds to a search where a user submit a query to a system and awaits a list of relevant documents among the entire collection. This task was evaluated during 8 consecutive years in TREC.

We chose this task because considering textual IR, it is one of the main tasks, which remained several years and for which there were a wide range of participants. From TREC adhoc data, we considered 7 years only (from 1993 to 1999); we preferred

to draw aside the first year which corresponded to the beginning of TREC framework and may not reflect the task itself.

Each year, TREC adhoc sends 50 queries to participants. In turn, each participant submits one or more runs. A run corresponds to the ranked list of the documents a system (and its parameters) retrieves for each of the 50 queries. The same research team participates to the task generally using only one system but under various configurations of parameters thus leading to several runs.

When considering the adhoc task, TREC server provides all the inputs files (ranked lists submitted to the National Institute of Standards and Technology, NIST) in a protected area. Table 1 indicates the number of runs that were sent to TREC adhoc and the corresponding topics.

Table 1 TREC collection features

TREC	TREC2	TREC3	TREC4	TREC5	TREC6	TREC7	TREC8
Topics	101-150	151-200	201-250	251-300	301-350	351-400	401-450
# systems	31	40	34	80	79	103	129

3.2 Evaluation

The bases of the evaluation in IR were developed within the Cranfield project [12]. A test collection includes a set of documents, a set of queries, and the list of the relevant documents for each query within the set of documents. The evaluation rests on the comparison between the list of documents that the system retrieves and the list of the relevant documents. Cleverdon puts forward two measures to evaluate the effectiveness of a search: the recall and the precision. Let us note that the Cranfield collection is composed of only 1400 documents; this size makes it possible to manually decide for each document whether it is relevant for a given query or not. Such an exhaustive evaluation is not possible when the collection is about one million documents, like in TREC or NTCIR (NII Test Collection for IR Systems, research.nii.ac.jp/ntcir). In these later cases, a pooling method is used in which the only documents that are judged are the documents that have been retrieved at least by one of the systems. The number of performance measures grew over years, but the majority of measures remains based on the assumption that the non-judged documents are non-relevant [31].

With regard to TREC, the performances of the systems are calculated by the trec_eval program written by Chris Buckley from Sabir Research [7]; it is also widely used by the IR community in general. This software includes various measures making it possible to evaluate the effectiveness of a system that produces a given run considering various criteria. The set of measures includes recall and precision at various levels of cut, as well as global measures derived from recall and precision. Recall measures the capacity of the system to find all the relevant documents; precision measures the capacity of a system to find only relevant documents. Whatever the measure, it is first calculated on each query; then the results are averaged over all queries with equal contributions. Let us note that the relevant documents are supposed to be known and non-judged documents are regarded as non-relevant. This potentially makes precision

smaller than it is since non-judged documents can be retrieved and be relevant; and recall generally overestimated since total relevant documents can be greater than the additional relevant retrieved documents.

Among 135 measures the version 8.1. of trec_eval (version from the 24th july 2006, visited on December 2007) evaluates, 5 are not associated with individual queries but with the average only. These measures have not been considered in this study: we kept only the 130 measures that can be associated to individual queries. All these measures are listed here
`www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README`

4 Statistical tools

4.1 Data handling

Multivariate statistical methods are generally based on a matrix of data as a starting point. From a statistical point of view, the considered matrix consists in rows, that correspond to individuals to analyze, and in columns, that correspond to variables used to characterize the individuals. In the present study, a row (an individual) corresponds to a run characterized by the performance measures, which indeed correspond to variables (columns).

Regarding individuals, depending on the year, the number of participants varies and the number of runs each participant sent varies as well. Theoretically, we should have had (31+40+34+80+79+103+129) systems * 50 queries = 24,800 rows (see table 1). However, some runs do not include all the 50 individual queries. We indicate in table 2 the number of rows we get for each TREC campaign.

Table 2 Number of rows per campaign in the analyzed matrix

TREC	TREC2	TREC3	TREC4	TREC5	TREC6	TREC7	TREC8
# rows	1550	2000	1618	3050	3700	5150	6450

In fact, the matrix we have analyzed is composed of 23,518 rows and 130 columns. An extract of the matrix we analyzed is presented in table 3.

In the present study, we considered the raw measures rather than their ranks because they are sharper than ranks. Intuitively, using ranks may pack down the differences between systems; but on the other hand it may expand the slight differences, and so can introduce some bias in measures. Moreover, in statistical studies, ranks are used to avoid outliers altering results of the analysis and to robustify it [22]. In our case, most of the measures are between 0 and 1; so outliers do not occur. Consecutively, the use of ranks is not relevant.

4.2 Sketch of the methodology

In our statistical approach, we focused on the columns of the matrix, in order to highlight the relationships between performance measures. To achieve this analysis,

Table 3 Extract of the analyzed matrix

Line	Year	System	Topic	0.20R.prec	0.40R.prec	0.60R.prec	...
1	TREC 1993	Brkly3	101	0.2500	0.1250	0.1111	...
2	TREC 1993	Brkly3	101	0.3077	0.2692	0.3077	...
3	TREC 1993	Brkly3	101	0.4737	0.4474	0.4211	...
...	...	...	...	...	...	...	...
23516	TREC 1999	weaver2	448	0.0000	0.0000	0.0357	...
23517	TREC 1999	weaver2	449	0.0000	0.0000	0.0000	...
23518	TREC 1999	weaver2	450	0.7627	0.6864	0.5966	...

we have considered three exploratory multivariate methods corresponding to the three stages of our approach:

- Stage 1: Analysis of the correlation matrix, which is the simplest method to consider the general correlation structure of performance measures;
- Stage 2: Clustering of the performance measures, in order to define a small number of clusters, each one including redundant measures; clusters defined in this way characterize the structure revealed in the first stage;
- Stage 3: Principal Component Analysis (PCA) that provides indicators and graphical displays giving a synthetic view, in small dimension, of the correlation structure of the columns and of clusters defined in the second stage. In the present study, PCA is mainly used to validate clusters obtained in the second stage.

These three methods are presented below. The reader used to perform multivariate statistical methods can skip the following subsections and directly go to section 5.

4.3 Correlation matrix

The correlation matrix analyzed here is a 130 $\times$ 130 symmetric matrix. Its general element, at the j-th row and the k-th column, is the linear correlation coefficient between the j-th and the k-th performance measures, generally called Pearson coefficient and defined as the ratio of the covariance over the product of the standard deviations. The value of this coefficient necessarily lies between -1 and +1. Its absolute value indicates the intensity, the importance, of this correlation: the closer to 1, the stronger the correlation. Additionally, the sign of the coefficient indicates the direction of the correlation: the two variables have a tendency to vary in the same direction when the sign is positive and in opposite directions when the sign is negative.

Unfortunately, if the number of variables is large (about a few dozens), the correlation matrix is rather difficult to read and to interpret. For this reason, more and more, these matrices are presented in a graphical form as in section 5.

4.4 Cluster analysis

General considerations Two types of clustering methods are usually used in statistics: agglomerative techniques (hierarchical clustering) and partitioning methods (for instance, k-means) [32, 33]. These two types are a little detailed below.

A clustering method can be performed indifferently on the rows or on the columns of the initial matrix. In some applications, it can even be relevant to carry out the two approaches. In the present work, the objective was to study the relationships between performance measures. The cluster analysis has been performed on the columns of the data matrix.

The majority of cluster analysis techniques require a measure of dissimilarity (for example, a distance) between each pair of objects (rows or columns) to classify. If the p variables of the data are homogeneous, the Euclidean distance (between rows or between columns) can be directly considered. Otherwise, when variables are heterogeneous, as a preliminary stage, each variable must be standardized (that is divided by its standard deviation). Note that when cluster analysis and PCA are successively performed on the same data, the two methods must be applied with the same Euclidean distance evaluated either with raw data or with standardized data.

Hierarchical clustering At the beginning of the algorithm, each variable constitutes a cluster (a class) composed of only this variable (thus there are p clusters). The algorithm starts then by gathering the two closest variables considering the chosen distance: after this stage, there are p-1 clusters. The algorithm continues then, gradually, to gather the two closest clusters, until obtaining a single cluster. This principle requires defining the distance between two clusters, which can be made in various ways, for example the distance between the average points of these clusters. Statisticians often advocate using the method known as the Ward method, which consists in fusing the two clusters that minimize the increase in the total within-cluster sum of squares [21]. This criterion, analogous to the inertia maximized in PCA, makes the two approaches coherent. The final result of such a hierarchical clustering method is a tree-like structure (or dendrogram) that can be cut at different levels in order to choose the number of clusters (or their size). The proportion of the total between-class sum of squares preserved by the partition obtained at a given level of cut is an interesting criterion to choose this level [21].

Partitioning methods These methods require two choices before beginning the algorithm: the number of clusters needed and a representative element for each cluster, called center. The first stage of the algorithm assigns each element to the nearest center; when all clusters are so defined, the algorithm recalculates the new centers, for example by taking the centroid (also called barycenter) of each cluster, and performs a new stage. The algorithm is finished when no element changes between two successive stages; this occurs in general very quickly (less than 5 iterations).

The approach considered here It is now admitted by statisticians that every time it is possible (when the number of objects to be classified should not exceed a few thousands) the best solution consists in performing successively the two methods, starting with the agglomerative one [22]. This approach has been followed here. In the first step, the Ward method was used with the Euclidean distance and the number of clusters had

been chosen examining the clustering-tree and the proportion of the total between-class sum of squares preserved at each level of cut (see section 5.1, *Clustering*). In the partitioning step, the initial centers were centroids of the chosen clusters. The distance between an element (an performance measure) and a center was the Euclidean distance. As a result of this step, the clusters from the hierachical method have been stabilized by reallocating elements at the border between two initial clusters.

4.5 Principal Components Analysis (PCA)

PCA is a very popular method in multidimensional statistical analysis. It makes it possible to produce graphical representations of the rows or of the columns of the considered matrix, and does so in a reduced dimension space. The method is defined so that the dispersion obtained in this reduced dimension space is the largest [19,23].

Let us consider the $(n \times p)$ data matrix denoted by X. The aim of PCA is to replace a p-dimensional observation (a row of X, here a run) by q linear combinations of the variables (here performance measures), where q (the dimension of the reduced space) is much smaller than p. The choice of these q linear combinations is made in order to explain a reasonable proportion of the trace of S ($\mathrm{tr}(S)$, the sum of the diagonal elements), a dispersion matrix. S can be the sample covariance matrix (when the p variables are homogeneous) or the sample correlation matrix (when they are heterogeneous). In order to obtain the greatest proportion of $\mathrm{tr}(S)$ with the smaller dimension q, the linear combinations defined by PCA are the eigenvectors related to the first q greatest eigenvalues. The ratio between each eigenvalue and $\mathrm{tr}(S)$ provide the proportion of the dispersion displayed by the related eigenvector. In practice, $q = 2$ or $q = 3$ are most often adequate. For more details on PCA see for instance [19,23].

Two kinds of graphical displays are possible as results of a PCA: a graphical display for rows of X, here individual queries from runs, and another one for columns of X, here performance measures. Only the latter one will be considered here since we focus on the relationships between these measures. Graphical display can be more meaningful by incorporating clustering results as a color of labels. We perform PCA in the third step in order to give a synthetic view of the results of the two previous methods used.

5 Results

In this section, we first analyse performance measures in order to define clusters of highly correlated measures (correlation analysis and clustering are used for this). Then we characterize the resulting clusters by presenting their content, analyzing their correlation using PCA, calculating the cluster homogeneity, and finally proposing some possible representative measure for each cluster.

5.1 Highly correlated measures

Correlation matrix In order to clarify the interpretation of the correlation matrix which includes 130 x 130 = 16900 values from which 8385 (130*129/2) are intrinsic, we represented it in the form of an image coding the numerical values between -1 and +1 according to different grey levels: the higher the correlation, the lighter the pixel.

Moreover, we proceeded to an optimal reorganization of the image produced in order to highlight more clearly the areas of strong correlation [10] (see figure 1). This figure is not to read pixel by pixel; it must be considered as the general picture of the strength of the correlations.

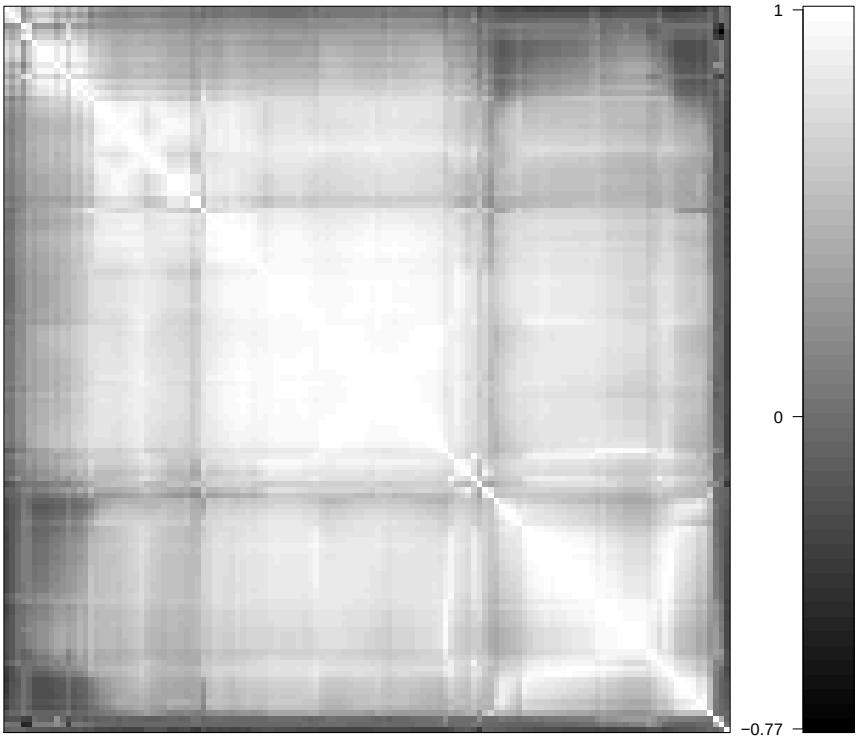

Fig. 1 Correlation matrix of the performance measures displayed using different grey level shades (from black -0.77 to white 1).

The fact that there are a lot of light patterns indicates a strong redundancy of information between performance measures: a "good" system is good whatever the performance measure used.

Numerically, 75% of the 8,385 intrinsic correlations values are greater than 0.45. This is a little lower than the 0.60 found on TREC7 only when considering a few measures [9], but note that a correlation calculated on more than 20,000 observations and greater than 0.45 is highly statistically significant.

However, some dark grey lines are disseminated through the image corresponding to negative correlations between some measures (583 correlation coefficients over 8,385 are negative). It concerns a small set of measures including rank_first_rel (rank of the first relevant document), num_ret (number of retrieved documents), num_nonrel_judged_ret (how many non-relevant documents have been retrieved) and others. For example for rank_first_rel, correlation with the 129 others measures vary from -0.37 to 1. Note that

for the large majority of the performance measures, the observed values are between 0 and 1 and a value close to 1 indicates that the system is good. On the other hand, for measurements rank_first_rel and num_nonrel_judged_ret it is not the case and interpretation can be: the smaller the measure, the better the system. Thus, the negative correlation of these latter measures with the others is expected. At last, num_ret is not actually a performance measure. The plots on figure 2 illustrate the behavior of some particular measures towards the others: the first two ones (rank_first_rel and num_ret) to highlight negative correlation, the two others (P5 and MAP) are selected as a reference to see the behavior of currently used measures.

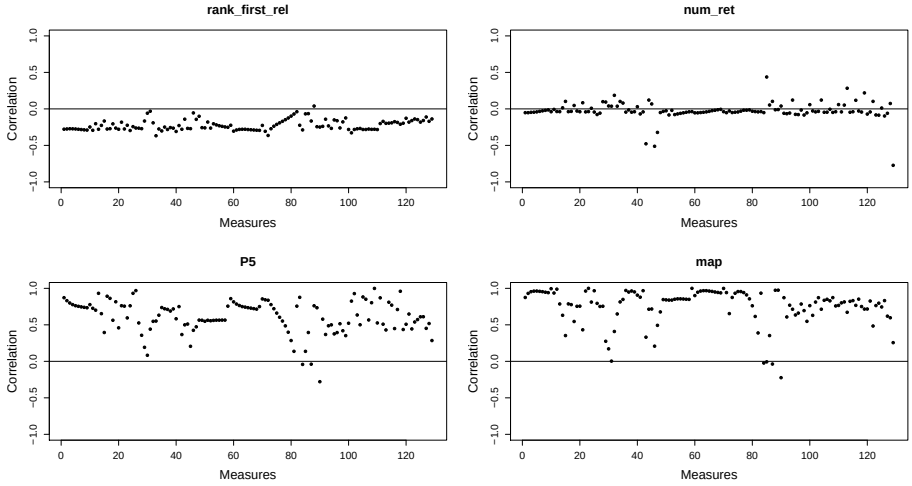

Fig. 2 For each measure rank_first_rel, num_ret, P5 and MAP, we plot the correlations between the measure concerned and the 129 others.

As it was shown in this section, the measures are highly redundant. As a consequence, a small number of performance measures must correctly summarize all of them. The aim of the clustering stage is to determine homogeneous clusters and then to define representative measures in each one.

The fact that some measures are not in the interval [0;1] leads us to convert original data into standardized variables (centered – mean 0 – and scaled – variance 1) to get rid of the problem involved by the various scale of several measures. Standardized measures are used in the sequel of this section. Note that the same type of reason justify normalizing average precision in [25].

Clustering After having highlighted obvious redundancies between performance measures, it is natural to consider a clustering process to limit the number of measures. Agglomerative clustering proposes a classification of performance measures without any prior information on the number of clusters. Here we have used the classical Euclidean distance and the Ward method.

The choice of the number of clusters is a crucial problem when performing clustering (see for instance [2, 11] in the context of text clustering). Our choice is based on the tree and the graph of the node heights (figure 3). The vertical scale of the tree represents

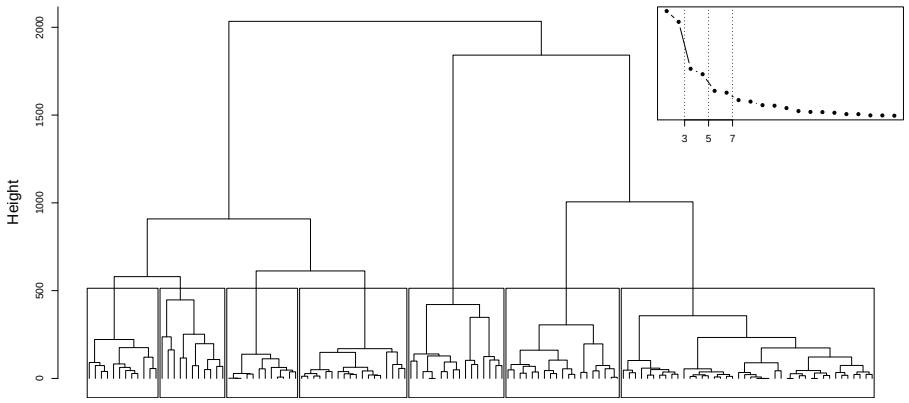

Fig. 3 Dendrogram representing the hierarchical clustering of the performance measures with a relevant pruning at 7 clusters. The sub-plot in the upper-right corner represents the height of the nodes of the dendrogram; it suggests relevant pruning at 3, 5 or 7 clusters.

the distance between the clusters; a relevant pruning level is characterized by a relative important difference between the heights of two successive nodes. On the sub-plot of figure 3, a relevant cut corresponds to a point for which there is a strong slope on the left and a weak slope on the right. Under these conditions, according to the degree of sharpness wished, one can retain here 3, 5 or 7 clusters.

Instead of using a cluster validation methods to determine an optimal number of cluster from hierarchical clustering, we opted for a heuristic consisting in selecting, among the reasonable choices, the highest number of clusters with a relevant meaning in terms of IR performance. Considering 7 clusters enable to distinguish 7 different behaviours that are detailed in the next section.

The results of hierarchical clustering were then stabilized using a k-means algorithm. This implied quite few changes: only 13 measures (among 130) concerning 3 clusters were reallocated.

Table 4 lists the measures in each of the 7 clusters obtained after the k-means method. The method we used aims at clustering measures that have the same behavior mathematically speaking; this does not mean that the clustered measures intend to measure the same retrieval property. Rather, the measures from a cluster tend to behave with the same trend. In addition, there are some retrieval properties that can be associated with each cluster.

5.2 Cluster contents

Cluster 1 (see table 4) is the second largest cluster and gathers 23 performance measures. It consists of the relative precision (precision after 100 documents relative to the maximum precision possible at that point; it considers each query to be equally important) for high precision (precision 5, 10, 15, 20 and 30), the relative unranked

Table 4 Clusters resulting from the successive use of hierarchical clustering and partitioning

Cluster 1 (23 measures)

relative_unranked_avg_prec30 - relative_unranked_avg_prec20 - relative_prec30 - map_at_R - relative_unranked_avg_prec15 - relative_prec20 - P30 - relative_prec15 - int_0.20R.prec - relative_unranked_avg_prec10 - X0.20R.prec - ircl_prn.0.10 - P20 - P15 - relative_prec10 - bpref_10 - P10 - relative_unranked_avg_prec5 - relative_prec5 - P5 - bpref_5 - recip_rank - ircl_prn.0.00

Cluster 2 (16 measures)

P100 - P200 - unranked_avg_prec500 - unranked_avg_prec1000 - bpref_num_ret - P500 - bpref_num_all - P1000 - num_rel_ret - exact_unranked_avg_prec - num_rel - exact_prec - bpref_num_correct - utility_1.0_.1.0_0.0_0.0 - exact_relative_unranked_avg_prec - bpref_num_possible

Cluster 3 (12 measures)

bpref_top10Rnonrel - bpref_retnonrel - relative_unranked_avg_prec500 - avg_relative_prec - recall500 - relative_prec500 - bpref_allnonrel - relative_unranked_avg_prec1000 - exact_recall - recall1000 - relative_prec1000 - exact_relative_prec

Cluster 4 (45 measures)

X1.20R.prec - ircl_prn.0.30 - X1.40R.prec - int_map - X1.00R.prec - R.prec - int_1.20R.prec - exact_int_R_rcl_prec - int_1.00R.prec infAP - avg_doc_prec - map - X11.pt_avg - X1.60R.prec - int_0.80R.prec - int_1.40R.prec - X0.80R.prec - old_bpref_top10pRnonrel - ircl_prn.0.40 - X1.80R.prec - int_1.60R.prec - X3.pt_avg bpref - X2.00R.prec - bpref_top25p2Rnonrel - old_bpref - bpref_top10pRnonrel - int_1.80R.prec - int_0.60R.prec - int_2.00R.prec - bpref_top25pRnonrel - X0.60R.prec - bpref_top50pRnonrel - bpref_top5Rnonrel - ircl_prn.0.20 - ircl_prn.0.50 - int_0.40R.prec - X0.40R.prec - int_map_at_R - ircl_prn.0.60 - unranked_avg_prec30 - ircl_prn.0.70 - ircl_prn.0.80 - unranked_avg_prec200 - unranked_avg_prec100

Cluster 5 (18 measures)

bpref_topnonrel - fallout_recall_42 - fallout_recall_28 - fallout_recall_56 - rcl_at_142_nonrel - fallout_recall_71 - fallout_recall_85 - relative_unranked_avg_prec100 - fallout_recall_99 - fallout_recall_113 - relative_prec100 - fallout_recall_127 - relative_unranked_avg_prec200 - fallout_recall_142 - recall100 - relative_prec200 - recall200 - bpref_retall

Cluster 6 (13 measures)

fallout_recall_14 - unranked_avg_prec20 - unranked_avg_prec15 - ircl_prn.0.90 - fallout_recall_0 - unranked_avg_prec10 - recall30 - ircl_prn.1.00 - recall20 - recall15 - unranked_avg_prec5 - recall10 - recall5

Cluster 7 (3 measures)

rank_first_rel - num_nonrel_judged_ret - num_ret

average precision 5, 10, 15, 20 and 30 (that is to say for high precision), the high precision it-self (precision P5 to P30), the interpolated recall (recall 0, 0.10), X0.20R.prec, int_0.20R_prec, MAP at R, bpref_5 and 10 and recip_rank. All the measures clustered into this group are high precision-oriented. This is the case for relative precision, relative unranked average precision, and precision at a certain number of retrieved documents; the measures when 5, 10, 15 and 20 documents are considered are grouped in this cluster. In addition, interpolated recall ircl_prn.0.00 and 0.10 implies low recall; since recall and precision vary oppositely, low recall implies generally high precision (apart for systems that are poorly performing in general). Moreover, these measures are computed when considering the first retrieved documents only. Considering a user's point of view, this group of measures is relevant to measure high precision; that is to say to measure the ability of systems to satisfy users who make use of the first retrieved documents. This usage is typical of users using search engines on the web [18]. More generally, this group of measures is relevant to measure the ability of systems to answer query for which a small set of documents is expected by users: this is the case for question/answering searches for which a single relevant answer is enough and searches that need a few answers. Systems that perform well considering the measures from this cluster do not retrieve necessarily all the relevant documents, but the first documents that are retrieved are relevant. These are general trends; however, the statistical point of view should be kept in mind: what the clustering shows is that even if each performance measure has been defined to measure a specific behavior of retrieval engines, systems behave the same way considering measures from one cluster, and this, even if the measured values are different. One measure from the cluster is enough as the representative measure of this cluster. How the representative measure is chosen is detailed section 5.3.

Oppositely, **cluster 2** groups together performance measures that consider large set of retrieved documents, if not the entire set. This is the case for precision measures when 100 or more documents are retrieved (P100 to P1000, exact_prec, exact_unranked_avg_prec). Considering a user's point of view, these measures are useful to measure the ability of systems to retrieve relevant documents whatever the size of the retrieved set of documents (even when retrieving a large set of documents). Typically, systems that perform well considering measures in cluster 2 are potentially filtering oriented [5] because these measures are less dependent to the order of the retrieved documents; they consider the set of retrieved documents rather than the ordered list of retrieved documents. This ability would satisfy users who want to gather as much as possible information on a given topic or field, or users who want to consider the different aspects of a query topic.

Cluster 3 consists of measures that are more recall-oriented (exact_recall, recall500, recall1000). Relative precision when a large number of relevant documents is considered are also clustered into cluster 3 (exact_relative_prec, relative_prec500 and 1000, avg_relative_prec, relative_unranked_avg_prec500 and 1000). Considering a user's point of view, such measures are related to the capacity of a system to retrieve most of the relevant documents. This is specifically interesting in tasks like science monitoring when sets of documents have to be gathered for further analysis or text mining [1].

The measure MAP, which is a global measure, is in a different cluster (**cluster 4**). It has been introduced in TREC2 because it aggregates recall/precision curves. It is thus not surprising to see that it is associated with ircl_prn.x (for x=0.20 to 0.8) in cluster 4. Ircl_prn.x for small values of x belongs to the high precision oriented measures whereas Ircl_prn.x for large values of x belongs to recall oriented measures.

Considering IR point of view, it also makes sense MAP and R-Prec (precision when considering a set of retrieved documents for which the cardinal is equal to the number of relevant documents) are grouped together. Indeed R-Prec can be seen as a measure of average precision since it does not imply to have a look only at the first retrieved documents nor the entire set of retrieved documents. This cluster contains also bpref and old_bpref, being introduced in 2004 and based on judged documents only. It is also a global measure of precision.

Measures Fallout_recall, recall100, recall200, relative_prec100 and relative_prec200 are grouped together in **Cluster5**. This cluster is more difficult to interpret in terms of IR properties and user-oriented features. It is composed of measures that are not heavily used. However, statistically speaking, the analysis shows that these measures tend to behave the same way when evaluating system performance.

Cluster 6 groups together interpolated recall ircl_prn.0.90 and .1.00 and recall at small cutoff levels (recall5 to recall30). Ircl_prn.90 measures the capability of a system to get good precision when most of the relevant documents have been retrieved. This measure tends to be close to 0 for a lot of systems (19291 values equal to 0 over the 23518 values for Ircl_prn.90). In the same way, specifically because the number of relevant documents is large, recall30 (recall when 30 documents are retrieved) is going to be small too. Considering TREC adhoc, these two measures get the same trend and thus are grouped together.

It can be noticed that **cluster 7** contains 3 particular measures already mentioned when analyzing the image of the correlation matrix. The use of k-means enabled to isolate these 3 measures from the others.

5.3 Cluster characterization

Principal Component Analysis To give a view of the relative proximities of the clusters, we performed a PCA restricted to the first 6 clusters (PCA performed on the 7 clusters essentially shows the specific behavior of cluster 7). The PCA brings another point of view on the relationships between the performance measures and confirms the redundancy of information as the major part of variability (67%) is supported by the first dimension PC1. The interpretation is focused on the dimension 1 and 2. Although it makes it possible to represent jointly variables (measures) and individuals (runs) of a data set, we focus here on the representation of the variables (figure 4). Each measure is represented by a symbol according to the cluster it belongs to.

The relative position of the clusters on the first and second principal components is consistent with the clusters obtained after the clustering process. Globally, the measures in each cluster appear projected relatively close to each other: figure 4 does not display a random mixing of symbols. Furthermore, it also offers a partial (because of the projection on a 2D space) representation of the inertia of the clusters. For instance, the measures of the cluster 3 (○), the most compact cluster according to the inertia criterion (figure 5), appear much closer to each other than the measures in other clusters.

Regarding PC1 (horizontal axis in figure 4), the main phenomenon is the opposition between clusters 3 (○), 5 (+) and 6 (■) on the left and, 1 (△) and 2 (▲) on the right. These relative positions highlight an opposition bewteen recall oriented clusters (3, 5 and 6) and precision oriented ones (1 and 2). Along PC2 (vertical), the opposition is between 1 (△) and 6 (■) (bottom) and, 2 (▲) and 3 (○) (top). In this case, the

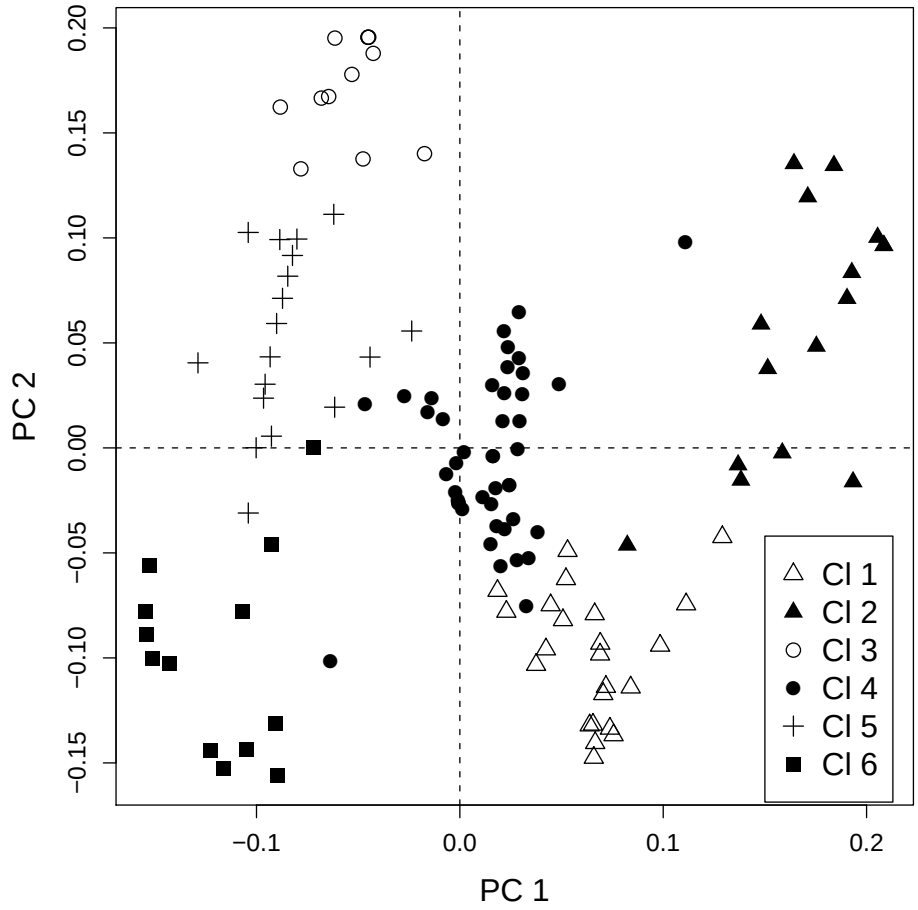

Fig. 4 Representation of variables on the first two principal components PC1 and PC2 respectively explaining 67% and 13% of the total variance. Symbols reveal the cluster the variables belong to.

discrimination globally concerns the number of documents on which is based the performance measure: few documents (less than 30) for clusters 1 and 6, and much more (more than 100) for clusters 2 and 3. Not suprisingly, cluster 4 (●), mainly composed of global measure aggregating recall/precision curves such as MAP, is located in the center of the plot. Cluster 4 act as an intermediate between other clusters.

Comparison of clusters homogeneity To provide a complementary interpretation of the clusters obtained after the clustering process, we calculate the within-cluster sum of squares for each cluster. As it highly depends on the number of points in each cluster, we used the ratio "within sum of squares / size of the cluster". We defined an inertia criterion by taking the square root of the ratio to be homogeneous as a standard deviation. The interpretation of such a criterion is: the smaller the value, the more compact the cluster. The figure 5 reveals that the most compact clusters are 1, 3, 4 and 5. For such a cluster, the choice of a representative element will be less important

as all the measures of the cluster are relatively close to each other. In other words, these clusters are so compact that any measure from one cluster can be a representative of this cluster (see table 4 as well). Indeed MAP for cluster 4 is a natural choice since it is widely used, as exact_recall could be a natural choice for cluster 3 for the same reason. P10 could be meaningful for cluster 1 since it corresponds to the cut-off most web search engines do when displaying results to users. Cluster 5 contains measures that are not as used by the community as the previous mentioned; a natural representative is not so obvious. On the contrary, clusters 2 and 6 are more heterogeneous and will be considered in the next sub-section.

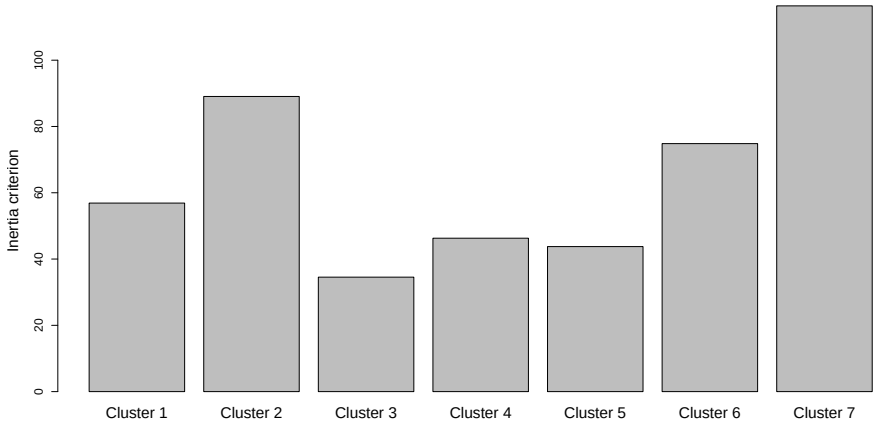

Fig. 5 Inertia criterion for each cluster.

Representative measures In this section, we aim at providing a method to define a small subset of measures summarizing nearly the same information as the 130 initial measures. As we previously pointed out, some retrieval methods are tuned in order to optimize a retrieval property that is relevant for the task the system is designed to, and thus to maximize the corresponding performance measure. However, it can be interesting to compare the results on the other measures. As it would be very fastidious to compare results on all the measures, it can be relevant to consider the minimal set of measures to use and avoid using two redundant measures.

To address this problem, we suggest choosing one (or more) representative in each cluster previously defined. The process we propose enables to identify potential candidates. It is important to understand that this process is one possibility among many.

Indeed, we show in the previous section that any measure from one cluster could be used. In the previous sub-section, we mentioned that, for compact clusters (1, 3, 4 and 5), mathematically, any measure is a good representative of the cluster and, for this reason, we suggested to choose among the most commonly used.

For less homogeneous clusters, we suggest the following general methodology composed of three steps:

– Calculate the centroid of each cluster by computing the arithmetic mean in each dimension.
– For every measure in each cluster, calculate the usual Euclidean distance to the centroid.
– Define the representative of the cluster as the nearest measure from the centroid.

This process is also an automatic alternative that can be used for the most compact clusters.

To identify the representatives, we display barplots (figure 6) representing, for each cluster (except the cluster 7 which is too singular to be analyzed as the others), the distance from the measure to the centroid. The values were sorted so that the most relevant measures are on the left of the charts (short bar).

In figure 6, for cluster 1, the barplot would suggest that P30 would be a slightly better choice than P10 to summarize the high precision measures. P100 makes sense for cluster 2. As we said previously, this cluster is less homogeneous than the previous clusters, but P100 is the closest to the centroid. Exact_recall is not the best choice for cluster 3, but given the compactness of this cluster, this choice is still mathematically relevant. The fourth barplot indicates that mathematically, MAP is a very good candidate for representing the cluster 4; bpref could be an alternative choice (same type of measures). Cluster 5 is the latest of the very compact clusters; any of the measures would make sense, bpref_topnonrel (or quasi-equivalently bpref_retall) would be, to our point of view, good candidates since they are easily understandable. Recall_30 could be an interesting candidate for cluster 6, specifically if P30 is chosen for cluster 1. Finally, rank_first_rel is the only element from cluster 7 that is really a performance measure.

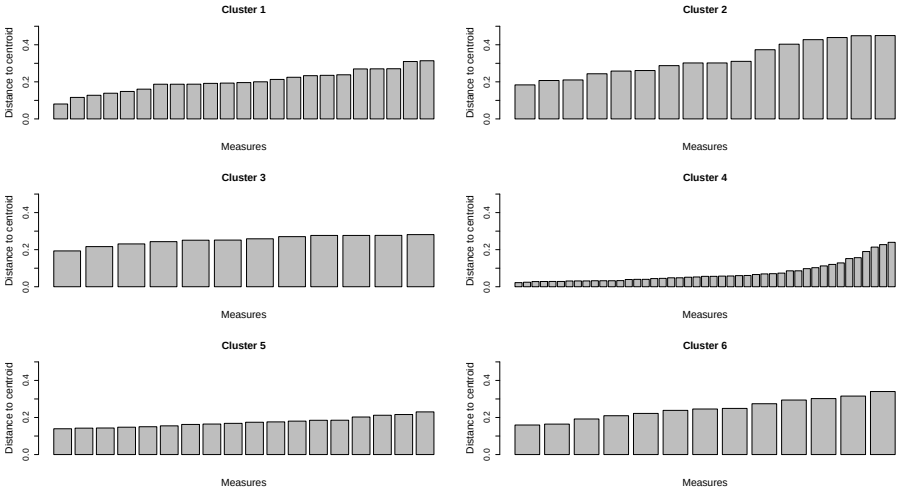

Fig. 6 Distance from the measures to the cluster centroid (each bar corresponds to a measure, the order they are drawn follows the one of table 4)

In the next section, we will focus on a subset of measures composed of: P30, P100, Exact_recall, MAP, bpref_retall, recall30.

6 Discussions

In this section, we made further analysis that intend to show the reliability of the analysis we made using large scale data in the previous sections. First, we compared the results obtain for the seven campaigns of TREC adhoc studied separately. Then we consider a laboratory framework in which the problem consists in ranking systems according to their performances.

6.1 One year analyses

The previous analysis is made globally on seven campaigns of TREC adhoc. This massive data set might hide some specificity of some years so we studied the seven collections separately. Detailed results are not shown here but we present global conclusions:

- Overall, there are just little differences in the clustering when considering a specific year. Mathematically the results are consistent from one year to another. About 90% of performance measures remained gathered in relation to the global study;
- The less stable clusters are the largest ones. Some of the differences correspond to different ways of grouping together measures of the same type. For example, in TREC4, int_0.20R.prec and ircl_prn.0.10 shift, but it corresponds to new clustering of ircl_prn.X and int.X.R.prec measures;
- For some years, the clustering process leads to eight clusters rather than seven. This is mainly due to the splitting of cluster 2 of the global study. This is consistent with the fact that, in figure 5, we identified this cluster as the less compact one (except the particular cluster 7). This implies a new cluster composed of exac_prec, exac_relative_unranked_avg_prec, exac_unranked_avg_prec, utility_1.0_.1.0_0.0_0.0.

These results strengthen the conclusions learned from the global analysis. The structure of the correlations of the performance measures are stable over year. So, the representatives previsouly defined will also be relevant for a less massive study.

6.2 Ranking systems

When considering laboratory experiments, the evaluation study is generally based on comparing several systems or, several versions of one system, and to decide which one is the best. In this section, we study two ranking processes based either on the complete set of performance measures or on the subset determined in section 5. The process is depicted in figure 7; it consists in ranking systems according to their average score. From the matrix that gathers, for one given topic, the value of each performance measure, we calculate an average score for each system, then the higher the average score, the lower the rank (and the better the system).

Detailed results are presented in an internal report [27]. We noticed globally that ranks do not change significatively when considering the complete set or the reduced set. Figure 8 illustrates this comparison for three topics. Scatterplots are very narrow indicating a nearly perfect relation between the two ranking processes. More precisely, very few changes occurs when looking at the best systems (smallest ranks) or the worth ones (largest ranks); this is consistent with the fact that systems are good, or bad,

$$\begin{pmatrix} x_{1,1} & \dots & x_{1,NbM} \\ \vdots & \ddots & \vdots \\ x_{NbS,1} & \dots & x_{Nbs,NbM} \end{pmatrix} \Longrightarrow \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_{NbS} \end{pmatrix} \Longrightarrow \begin{pmatrix} r_1 \\ \vdots \\ r_{NbS} \end{pmatrix}$$

Fig. 7 Schematic view of the process from scores to ranks. NbS and NbM respectively denotes the number of systems and the number of performance measures. $Nbm = 118$ or 6 when considering either the complete set (except measures whose score is not in [0;1]), or the reduced set considering a representative measure in each cluster except cluster 7.

whatever the evaluation used. In addition the Pearson correlation coefficient computed on average scores from the complete set and the reduced set was systematically around 0.99.

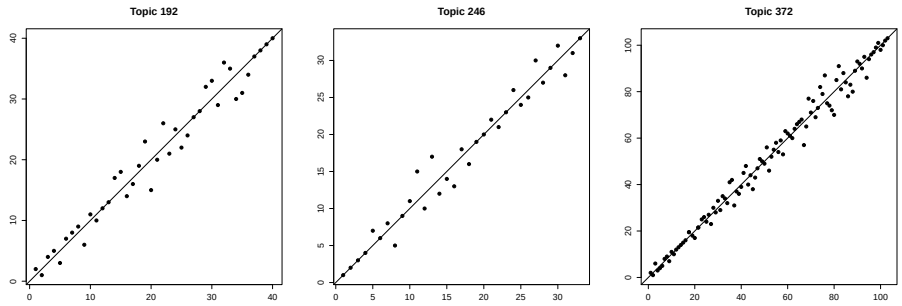

Fig. 8 Comparison of system ranks based on the complete set (horizontal) or the subset (vertical) of performance measures for three topics: #192 from TREC3 (40 systems), #246 from TREC4 (34 systems) and #372 from TREC7 (103 systems).

7 Conclusions

In this paper, we analyzed massive sets of TREC adhoc results. We considered the official runs submitted each year to this task, each consisting in 50 lists of retrieved documents (50 topics are used each year). We then considered the 130 measures provided by trec_eval for individual queries. The matrix we analyzed is then composed of 130 columns and 23,518 rows.

The objective of this paper was to group together measures into coherent clusters. The methodology we applied combines different statistical methods. The main results of this study are the following:

- We show that all the performance measures are highly correlated when considering TREC adhoc data.
- We confirm thus the trend that good systems are good, whatever the measure used; meaning that a system which performs well considering a measure will tend to perform well considering a different measure.

- We show that even if performance measures have been introduced to measure some retrieval property, they can be grouped together into coherent families. That does not mean that they measure the same thing but measured clustered together tend to behave the same way.
- We extract seven highly correlated clusters of measures from past TREC adhoc results. For each family we indicate how much they are correlated. For the most homogeneous clusters, any measure is a good representative of the cluster from the mathematical point of view. For the other clusters, we suggest a method based on the centroid and consider both the distance to the centroid and the popularity of the measure usage.
- We characterize some of the clusters not only on a mathematical point of view but on an information retrieval point of view as well.

The analysis of the results per year yields the same conclusions as mentioned above showing that the results are not collection dependent.

Even if retrieval systems are usually design to perform a given task, it can be useful to compare two systems or two tuning of the same system on different measures. Intuitively, search engine designers know that s/he should not consider one single value. Tuning a system to improve P5 only for example might be a too slight view since even other high precision measures might be bad, not speaking of more recall oriented measures. We show that one measure from each cluster is enough to get a full picture.

As an additional result of the paper, we promote a method to analyze any new measure. New measures can be introduced in order to measure one retrieval property or another previous measures were not defined for. For example, some measures have been recently added such as bpref [8] in order to overcome some bias related to the fact that relevance is based on incomplete sets [31]. More recently, Mizzaro and Robertson [25] define normalized versions of average precision. Applying the analysis method we developed in this paper could help when defining new measures, not to know if they measure a new retrieval characteristic, but rather whether or not it is redundant in terms of behavior with existing measures and thus if the phenomenon to study is already measured.

This paper promotes the use of data-mining techniques that can lead to many new tracks in IR [35]. New developments in topic categorization [28], hard topic detection [26] and system fusion [4] are some of the examples.

Acknowledgements This work has been supported by Programme Pluri-Formations de Recherche en Mathématiques et Informatique de Toulouse (PPF FREMIT) Université Paul Sabatier, Toulouse III.

References

1. Alaux, J., Dousset, B., Chrisment, C., Mothe, J. (2003). DocCube: Multi-Dimensional Visualisation and Exploration of Large Document Sets, Journal of the American Society for Information Science and Technology, 54(7):650-659.
2. Al Hasan, M., Salem, S. Zaki, M.J. (2010). SimClus: an effective algorithm for clustering with a lower bound on similarity. Knowledge and Information Systems, DOI 10.1007/s10115-010-0360-6 (accepted oct. 2010).
3. Aslam J. A., Yilmaz E., Pavlu V. (2005). A geometric interpretation of r-precision and its correlation with average precision, Proceedings of the international ACM SIGIR conference on research and development in information retrieval, 573-574.

4. Bigot A., Chrisment C., Dkaki T., Hubert G., Mothe J. (submitted 2010 IRJ), Fusing different information retrieval systems according to query topics a study based on correlation in information retrieval systems and query topics (will be removed if rejected, revised version submitted in December 2010).
5. Belkin, N. J., Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin?, Communication of the ACM, 35(12):29-38.
6. Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems, Information Research, 8(3), paper no. 152 [Available at: http://informationr.net/ir/8-3/paper152.html].
7. Buckley, C. (1991). Trec_eval,available at `http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec\_eval\_video/README`
8. Buckley, C., Voorhees, E. M. (2004). Retrieval evaluation with incomplete information, Proceedings of the international ACM SIGIR conference on research and development in information retrieval, pp 25-32.
9. Buckley C., Voorhees, E.M. (2005). Retrieval system evaluation. In Voorhees E.M. et Harman D.K., TREC : experiment and evaluation in information retrieval, MIT Press, 53-75.
10. Caraux, G., Pinloche, S. (2005). Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order, Bioinformatics, 21, 1280-1281.
11. Chen, C.-L., Tseng, F.S.C., Liang, T. (2010). An integration of fuzzy association rules and WordNet for document clustering, Knowledge and Information Systems, DOI 10.1007/s10115-010-0364-2 (accepted nov. 2010).
12. Cleverdon, C. W., Mills, J., Keen, E. M. (1966). Factors determining the performance of indexing systems. Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics, (Volume 1:Design; Volume 2: Results).
13. Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations, Information Processing & Management, 44(2):856-876.
14. Hersh, W.R., Elliot, D.L., Hickam, D.H., Wolf, S.L., Molnar, A., Leichtenstien, C. (1994). Towards new measures of information retrieval evaluation, Proceedings of the annual symposium computer application in medical care, 895-899.
15. Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments, Proceedings of the international ACM SIGIR conference on research and development in information retrieval, 329 338.
16. Ishioka, T. (2003). Evaluation of criteria for information retrieval, Web Intelligence, WI 2003, Proceedings IEEE/WIC International Conference, 425- 431.
17. Jarvelin, K., Keklinen, J. (2000). IR evaluation methods for retrieving highly relevant documents, Proceedings of the international ACM SIGIR conference on research and development in information retrieval, 41-48.
18. Jansen, B. J., Spink, A., Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web, Information Processing & Management, 36(2):207-227.
19. Jolliffe, I.T. (2002). Principal Component Analysis, second edition, Springer.
20. Kurland, O. (2009). Re-ranking search results using language models of query-specific clusters, Information Retrieval Journal, 12(4):437-460.
21. Lebart, L., Morineau, A., Warwick, K.M. (1984). Multivariate Descriptive Statistical Analysis, Wiley.
22. Lebart, L., Piron, M., Morineau, A. (2006). Statistique exploratoire multidimensionnelle : Visualisations et inférences en fouille de données, 2006 , 4ème édition, Dunod.
23. Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). Multivariate Analysis, Academic Press.
24. Melucci, M. (2007). On rank correlation in information retrieval evaluation. ACM SIGIR Forum, 41(1), 18-33.
25. Mizzaro, S., Robertson, S. (2007). Exploring IR Evaluation Results with Network Analysis, Proceedings of the international ACM SIGIR conference on research and development in information retrieval, 479 486.
26. Mothe J., Tanguy L. (2005). Linguistic features to predict query difficulty - A case study on previous TREC campaign, SIGIR workshop on Predicting Query Difficulty - Methods and Applications, 7-10.
27. Poirier J. and Sansas B. (2009). Comparaison des classements de systèmes de recherche d'information en fonction des mesures de performances utilisées [Comparing IRS ranks in function of the evaluation measures that are used]. Internal Report NIRIT/RR–2009-31–FR, IRIT.

28. Pu, H-T., Chuang, S-L,, Yang, C., (2002). Subject categorization of query terms for exploring Web users' search interests, Journal of the American Society for Information Science and Technology archive, 53(8): 617-630.
29. Robertson, S.E. (1981). The methodology of information retrieval experiment. In: Sparck Jones, K. ed., Information retrieval experiments. London: Butterworths, 9-31.
30. Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance, Information Processing and Management, 43(2):531-548.
31. Sakai T., Kando N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments, Information Retrieval Journal, 11(5): 447-470.
32. Sakuma J. and Kobayashi S. (2010). Large-scale k-means clustering with user-centric privacy-preservation. Knowledge and Information Systems, 25(2): 253-279.
33. Seber, G.A.F. (1984). Multivariate Observations, Wiley.
34. Tague-Sutcliffe J. and Blustein J. (1995). A statistical analysis of the TREC3 data. In Proceedings of the Third Text REtrieval Conference (TREC-3), 385-398.
35. Taniar, D., (2007). Research and Trends in Data Mining Technologies and Applications, Information Retrieval Journal, 11(2):165-167.
36. Voorhees E.M., Harman D. (1999). Overview of the Eighth Text REtrieval Conference (TREC-8) Proceedings NIST Special Publication:SP 500-246, 1-23.
37. Voorhees, E.M. (2002). The Philosophy of Information Retrieval Evaluation, Lecture Notes in Computer Science, Volume 2406/2002, ISSN 0302-9743, Springer Berlin / Heidelberg.
38. Voorhees, E.M. (2007). Overview of the TREC 2006, The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings NIST Special Publication:SP 500-272, 1-16.
39. Webber, W., Moffat, A., Zobel, J., and Sakai, T. (2008). Precision-at-ten considered redundant. Proceedings of the international ACM SIGIR conference on research and development in information retrieval, 695-696.
40. Yilmaz, E., Robertson, S. (2010). On the choice of effectiveness measures for learning to rank, Information Retrieval Journal, Special issue on Learning to rank for information retrieval, 13(3): 271-290. 10.1007/s10791-009-9116-x.