

A New Evaluation Measure for Information Retrieval Systems

Martin Mehlitz
Technical University Berlin, DAI-Labor
10587 Berlin, Germany
martin.mehlitz@dai-labor.de

Christian Bauckhage
Deutsche Telekom Laboratories
10587 Berlin, Germany
christian.bauckhage@telekom.de

Jérôme Kunegis
Technical University Berlin, DAI-Labor
10587 Berlin, Germany
jerome.kunegis@dai-labor.de

Şahin Albayrak
Technical University Berlin, DAI-Labor
10587 Berlin, Germany
sahin.albayrak@dai-labor.de

Abstract—Some of the established approaches to evaluating text clustering algorithms for information retrieval show theoretical flaws. In this paper, we analyze these flaws and introduce a new evaluation measure to overcome them. Based on a simple yet rigorous mathematical analysis of the effect of certain parameters in cluster based retrieval, we show that certain conclusions drawn in the recent literature must be taken with a grain of salt. Our new measure, in contrast, accounts for statistical biases that have to be expected according to our analysis. A series of experiments and a comparison with results reported recently underlines that this measure is a more suitable performance indicator that allows for more meaningful interpretations.

I. INTRODUCTION

Information retrieval is the discipline within computer science that deals with the automated storage and retrieval of documents and data [1], [2], [3], [4]. As the amount of data available to modern information societies continues to grow rapidly, it is not surprising to find that information retrieval is a very active area of research.

In this paper, we focus on the problems of text retrieval and text clustering where the latter has been proposed as a technique for improving text retrieval [5]. In general, systems that apply text clustering either retrieve documents by automatically ranking clusters and then selecting documents from the highest ranking clusters [6] or by showing the cluster structure to the user who has to select a cluster whose documents are then displayed [7]. The output of a system that automatically ranks clusters is a flat list of documents similar to that of a regular information retrieval system. Therefore, the evaluation of these systems is similar to the evaluation of regular information retrieval systems. However, evaluating a system that shows a cluster structure to a user though requires other measures. A methodology that is often applied for evaluation of these systems is the *optimal cluster search* which selects the cluster with the optimal value in regard to the employed measure [8].

In this paper, we analyze the pitfalls of this evaluation methodology. Based on an appropriate mathematical model for what is going on in cluster-based text retrieval, we reconsider results reported in a recent paper which evaluates

several cluster algorithms for query-specific clustering [9]. Our main contribution will be a new measure for evaluating information retrieval systems, which circumvents the shortcomings which we identify in our analysis.

The remainder of this paper is structured as follows: In Section II, we will briefly review concepts applied in information retrieval. Section III will then present a mathematical analysis of how the optimal cluster search methodology affects precision and recall measures. Based on the results we derived for random clusterings, Section IV will introduce a new evaluation indicator, which, in the experiments in Section V, is compared to the classic evaluation measures. A summary and an outlook on future work will conclude this contribution.

II. BACKGROUND AND RELATED WORK

A. Clustering Documents for Information Retrieval

Clustering is the process of grouping items in a way that items in a group are similar to each other and dissimilar to the items in other groups [10].

In information retrieval often the problem for retrieving the desired documents given a short query lies in the ambiguity of the query. Terms in a query can occur in various documents and contexts and usually it is very difficult to identify the documents where a search term is used in the intended sense. Therefore, many search results in a result list of a classic information retrieval system are irrelevant. Clustering results for information retrieval is a way to circumvent this problem by structuring the (probably) very long list of search results. Of course, this will not avoid any irrelevant documents but if the cluster labels are intuitive enough, the user can directly choose the cluster which contains the most relevant documents for his request.

Figure 1 illustrates the process of document clustering. Starting with a user querying an information retrieval system with the phrase “jaguar”, a list of relevant documents is found by matching strings in the search phrase with documents. While most search engines stop after the second step (the actual retrieval of documents), text clustering proceeds by analyzing the documents in the result list in order to group

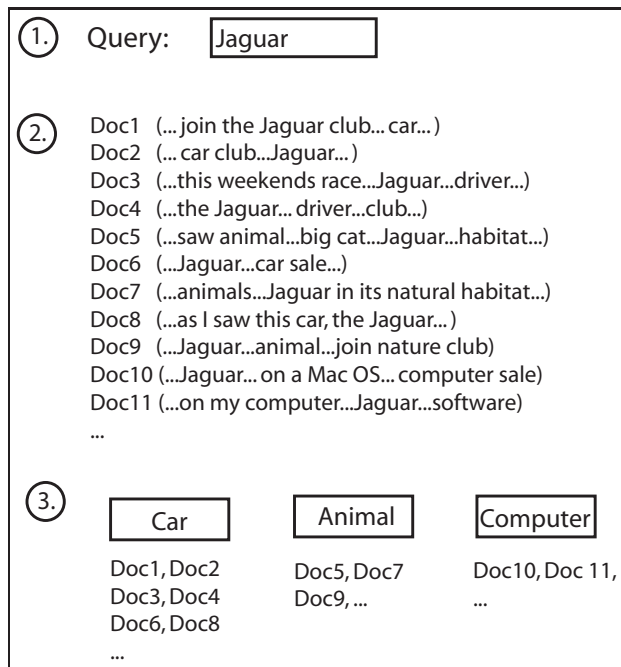


Fig. 1. Didactic example of clustering documents for information retrieval

similar documents. The user who queried for “jaguar” can then choose among the clusters and access the corresponding documents.

B. Evaluation of Information Retrieval Systems

The measures typically employed to evaluate an information retrieval system are *precision* and *recall*. *Precision* (“How many of the retrieved results are relevant?”) and *recall* (“How many of the relevant results are retrieved?”) are defined as:

$$P(g_{ret}, r) = \frac{g_{ret}}{r} \quad (1)$$

and

$$R(g_{ret}, g) = \frac{g_{ret}}{g}, \quad (2)$$

where r denotes the number of documents retrieved, g corresponds to the number of relevant documents in the collection and g_{ret} denotes the number of relevant documents among the returned documents. The effectiveness measure E is proposed in [8] as a measure for the evaluation of information retrieval systems that are based on text clustering. E is defined as:

$$E(P, R) = 1 - \frac{(\beta^2 + 1) PR}{\beta^2 P + R}, \quad (3)$$

where β is a factor that determines the relative importance of *precision* and *recall* where *precision* and *recall* values are based on the documents in the cluster.

C. Evaluation of Document Cluster Information Retrieval Systems

Based on the hypothesis that “*closely associated documents tend to be relevant to the same request*” [4] – some

information retrieval systems employ document clustering in order to achieve improvement in retrieval of relevant documents. Especially hierarchical cluster methods are very popular for text clustering, since hierarchies of clusters are supposed to be relatively intuitive to browse. Research on using document clustering techniques for information retrieval resulted in algorithms like Suffix Tree Clustering [7] and the DisCover Algorithm [11] as well as in the design of the Scatter/Gather System [12] and the CREDO System [13].

Usually, novel document clustering methods are compared to existing methods or to an inverted file search, based on *recall* and *precision* measures or some variants of these. A common evaluation methodology for such a comparison is the “*Optimal Cluster Search*” [7], [8], [9], [12], [14]. This methodology evaluates every single cluster and then selects the cluster with the optimal value as a representative for the cluster hierarchy.

Text clustering for information retrieval can be done either statically on the whole collection of documents [6], [8] or in a query-specific manner for a small number of top ranking documents returned by an inverted file search [7], [13]. In [9], Tombros, Villa and van Rijsbergen evaluate several cluster algorithms on different data sets with the goal of identifying the extent to which the number n of top ranking documents that are clustered influences the retrieval of relevant documents.

When creating document clusters, documents can either be limited to occur in only one leaf cluster of a hierarchy (which means they still appear in their parent clusters all the way to the root) or they are allowed to occur in multiple clusters. Note that forcing documents to appear in only one cluster is considered to be artificial [7], [12] since documents can contain more than one topic and a good cluster structure should reflect this.

III. COMPUTING EXPECTED VALUES FOR RANDOM CLUSTERING

When evaluating a text clustering algorithm for information retrieval with the optimal cluster evaluation methodology one has to consider the impact of this methodology on the achieved results. If an algorithm produces a large number of clusters and the best one is selected, then the critical point is, how likely it is to achieve such a result if the clusters were generated randomly. In this section we review the mathematical background for computing the expectation of random clusterings. This expectation value can indicate whether the outcome of an experiment was likely to happen by chance or not.

First, we explain how to compute the value for a random pick of pages from a set of documents. Then we address how a growing number of random picks influences this value.

A. One Random Pick

The probability distribution for the random pick scenario is a hypergeometric distribution: Given a finite set of items (documents) where some are good (relevant) and some are bad (irrelevant), we draw without replacement. Suppose we

have a set of items consisting of g good items and b bad items. If we draw exactly d items, the probability for drawing a number i of good items is given by:

$$P(x = i) = \frac{\binom{g}{i} \binom{b}{d-i}}{\binom{g+b}{d}}. \quad (4)$$

The corresponding expected value for this probability amounts to:

$$\mu(d) = \sum_{i=0}^d \left(i \frac{\binom{g}{i} \binom{b}{d-i}}{\binom{g+b}{d}} \right) = \frac{g}{g+b} d. \quad (5)$$

Writing this expectation as a function of d , the total number of items drawn from the set of available items, shall emphasize that, if g and b are fixed, it only depends on how many items are drawn.

B. Multiple Random Picks

The next step is to analyze to which extent the methodology of first generating multiple clusters and then selecting the best one changes the expected value of good items. Therefore, we need to express the likelihood of an outcome of a certain number of occurrences of relevant pages with regard to the number and size of clusters. Suppose we create c clusters where each cluster contains d items. Since the optimal cluster methodology selects the best cluster from a set of clusters, we need to describe the probability that a number of i relevant pages is the maximum number of relevant pages in the set of c clusters and that i really occurs. It is given by:

$$P(x = i, c, d) = \frac{B(x = i, d, c)}{A(d, c)}, \quad (6)$$

where A is the total number of different outcomes of choosing c times the number of d items and where B is the number of different ways of choosing these items with i being the maximum number of good items in the set of c clusters and with i really occurring. A is given by:

$$A(d, c) = \binom{g+b}{d}^c, \quad (7)$$

while B is given by:

$$B(x = i, d, c) = \sum_{j=1}^c \binom{c}{j} C(i, d)^j D(i, d)^{c-j}, \quad (8)$$

where j denotes the number of clusters that actually contain i good items. The first factor in the sum denotes the number of ways of distributing these j good items over the clusters c . C is the number of ways of drawing the i good items, while D counts the number of ways of drawing less than i good items. C and D are given by:

$$C(i, d) = \binom{g}{i} \binom{b}{d-i} \quad (9)$$

and

$$D(i, d) = \sum_{k=0}^{i-1} \binom{g}{k} \binom{b}{d-k} \quad (10)$$

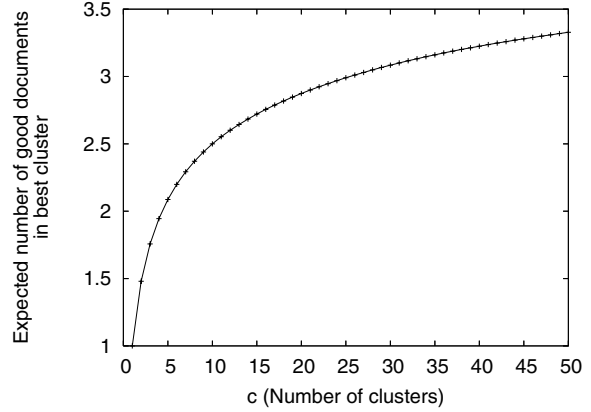


Fig. 2. Expectation values obtained from applying optimal search to randomly generated document clusters.

respectively.

The expectation for multiple random draws is given by:

$$\mu_{all}(d, c) = \sum_{x=0}^d x P(x, c, d) \quad (11)$$

which only depends on the number of clusters and the number of pages in each cluster.

C. A Random Cluster Example

Given the result in (11), let us consider a simple example of how the optimal cluster evaluation method can distort the evaluation results for a information retrieval. Suppose there are a 100 documents, 10 of which are relevant. From these documents we always draw ten documents. Fig. 2 shows the expected number of relevant documents that the best draw (optimal cluster) contains in relation to the number of times that we draw. Note that a small number of clusters already yields quite a large number of relevant documents: Selecting the best cluster out of five random clusters yields an expected number of relevant documents that is more than twice as high than that of a single draw. Already this simple experiment underlines how important it is either to report detailed information on the experimental statistics (which unfortunately is not common practice) or to use an evaluation measure which incorporates them.

IV. A NEW MEASURE FOR EVALUATION OF INFORMATION RETRIEVAL SYSTEMS

Recall and *precision* have been widely employed by the information retrieval community in order to evaluate information retrieval systems. We propose new measures called *absolute precision* and *absolute recall* that incorporate the knowledge of the expected value of a random retrieval with similar settings for a given scenario and experiment. We define them as:

$$P_{abs}(g_{ret}, g_{exp}, r) = \frac{g_{ret} - g_{exp}}{r} \quad (12)$$

$$R_{abs}(g_{ret}, g_{exp}, g) = \frac{g_{ret} - g_{exp}}{g} \quad (13)$$

With g_{exp} , the number of relevant documents that can be expected to be retrieved by a random retrieval.

Therefore, if the information retrieval system does not employ document clustering, the value of g_{exp} can be computed by (5). For a system that does employ document clustering methods, (11) can be used to compute this value if the optimal cluster search is used for evaluation. These new values for *precision* and *recall* can then be used to compute the *absolute effectiveness* by slightly adjusting (3):

$$E_{abs} = E(P_{abs}, R_{abs}) \quad (14)$$

Other measures based on *recall* and *precision* can be derived in a similar fashion. Note that P_{abs} and R_{abs} can be negative given that a random retrieval performed better than an actual experiment. Therefore, the values of E_{abs} will be larger than 1 for these cases.

V. EXPERIMENTS

In this section we will employ the theoretical model for computing expected values for clustering documents and compare them to the results reported by Tombros et al. [9] by using the new evaluation measure introduced above.

A. Experimental Settings

There are several document collections and cluster algorithms evaluated by Tombros et al., but for simplicity we shall consider only one collection and one algorithm. Since statistics such as average cluster size are not being reported for every combination of algorithms and document collection, we will focus on the WSJ text corpora¹ and the group average cluster algorithm for which these statistics are most complete. The group average cluster algorithm is also the one which performed best in [9].

For computing the expected values of an optimal cluster evaluation, we need to know the number of documents n that are clustered and the number of relevant documents g among these documents. Furthermore, we need to know the average cluster size d and the average number of generated clusters c . Unfortunately, Tombros et al. only report the first three of these values while the average number of generated clusters is missing. Only for their experiments with the full LISA collection² (6004 documents) do they report the number of clusters (6003). In order to be able to compute the expected values for the experiments we assume a relatively low number of generated clusters for each value of n .

Since we need integer values for g and d , we round both values down to the next smaller integer. Since the expectation in (11) is growing monotonously in these values, this will result in values that are slightly smaller than the real expected values for the experiments, favoring the results in [9]. The actual values for the settings in each experiment are shown in Table I.

¹The Wall Street Journal (WSJ) text collection is part of the Tipster corpus which can be obtained from <http://trec.nist.gov/>

²A collection of abstracts of the *Library and Information Science Abstracts* from 1982

TABLE I
EXPERIMENTAL SETTINGS

scenario number	number retrieved	number relevant	generated clusters	average size
1	100	16	50	12
2	200	24	100	17
3	350	31	175	21
4	500	37	250	24
5	750	43	375	28
6	1000	47	500	31

B. Experimental Results

With the above settings we computed the expected number of relevant documents in the optimal cluster for a random experiment. These values are used to compute the *absolute effectiveness*. For each scenario, we give the effectiveness values (the smaller the better) reported by Tombros et al. as well as the values for the *absolute effectiveness*. Table II compares the values reported in their experiments to the values that can be expected with randomly generated clusters. Note that for small values of n even though the effectiveness reported is relatively low the *absolute effectiveness* is not. In fact, the *absolute effectiveness* for the value of $n = 100$ top ranking documents is higher than one, indicating that randomly generating a similar number of cluster with a similar size yields better results than the results reported by Tombros et al.

One of the goals of Tombros et al. was to determine the importance of the parameter n for the retrieval. They conclude that this number has very little impact on the achieved effectiveness. However, according to our experiments it appears that one should avoid making any statement about the results achieved with small values of n , since the employed cluster algorithms did not manage to produce a significantly good cluster structures.

There are several issues concerning the question whether this comparison is fair or not. First, a source of error in the computation of random clusters is that we take the average cluster size to compute expected values for relevant documents. Therefore, for setting the parameter $\beta \neq 1$ in the expression used for computing effectiveness (see (3)), the effectiveness reported in [9] will always show a positive bias, since the size of optimal clusters changes with this parameter (as acknowledged by the authors).

Furthermore, we assume a hypergeometric distribution of documents in each cluster. This basically applies to cluster algorithms that allow documents to be included in more than one cluster at the same level of the hierarchy. The algorithms evaluated by Tombros et al. though are agglomerative hierarchical cluster algorithms, therefore a document is only included in one leaf node. However, when Tombros et al. evaluate cluster hierarchies they do not limit their search to leaf nodes but also regard intermediate nodes, therefore documents actually do occur in more than one node in the evaluated hierarchy.

TABLE II
EXPERIMENTAL RESULTS

scenario number	expected number of relevant documents	reported effectiveness			absolute effectiveness
		$\beta = 0.5$	$\beta = 1.0$	$\beta = 2.0$	$\beta = 1.0$
1	4.85	0.608	0.692	0.696	1.038
2	5.47	0.604	0.67	0.661	0.943
3	5.85	0.603	0.671	0.65	0.896
4	5.99	0.585	0.668	0.642	0.864
5	5.98	0.585	0.667	0.64	0.835
6	5.88	0.586	0.676	0.641	0.827

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we have shown that the commonly used methodology for evaluating the *optimal cluster search* approach to Information Retrieval may result in misleading interpretations. Still, we consider evaluating every single cluster and selecting the one with the optimal score to be a valid tool for examining clustering approaches to retrieval, if, that is, one takes into account that certain observations are basically due to the evaluation methodology. For instance, in a recent survey on the effectiveness of clustering-based retrieval algorithms [9], the authors report that they did not find a statistically significant variation in query-specific cluster effectiveness for different values of top-ranked documents.

From the rigorous theoretical analysis presented in this paper, however, we can conclude that the algorithms they evaluated had to perform poorly given the experimental setting and parameterization that was considered. Concerning examples like this, our main contribution in this paper is therefore the definition of a new measure for evaluating information retrieval systems that accounts for statistically expectable results by normalizing the performance measure with respect to these. In a series of experiments that analyze the impact of the number of clustered documents on the cluster effectiveness, we have shown that the use of this new measure yields meaningful results. Compared to the figures obtained from conventional precision and recall measures the absolute precision and absolute recall indicate, for instance, that no serious statement can be made for small numbers of clustered documents.

Since the influence of the number of top ranking documents that are clustered on the effectiveness of the clustering was not sufficiently analyzed by [9], we will further investigate this influence based on the more appropriate performance measures which we introduced in this paper.

Machine learning techniques that are employed for learning the parameters of a text cluster algorithm depend on an

error function that indicates “goodness” of a certain set of parameters. We will compare the training results achieved by error functions based on conventional performance indicators to those achieved with error functions based on evaluation measures derived from our model of expectation values for random clustering.

REFERENCES

- [1] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, Kluwer, Norwell, MA, 1997.
- [2] M.T. Maybury, Ed., *Intelligent Multimedia Information Retrieval*, MIT Press, Cambridge, MA, 1997.
- [3] G. Salton, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY, 1983.
- [4] C.J. van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, 1979.
- [5] G. Salton, *Automatic Information Organization and Retrieval*, McGraw Hill, New York, US, 1968.
- [6] W.B. Croft and C.J. van Rijsbergen, “Document clustering: An evaluation of some experiments with the cranfield 1400 collection,” *Information Storage Retrieval*, 11, pp. 171–182, 1975.
- [7] O. Zamir and O. Etzioni, “Web document clustering: A feasibility demonstration,” in *SIGIR*, 1998, pp. 46–54, ACM.
- [8] N. Jardine and C.J. Van Rijsbergen, “The use of hierarchic clustering in information retrieval,” *Information Storage and Retrieval*, vol. 7, pp. 217–240, 1971.
- [9] A. Tombros, R. Villa, and C. J. van Rijsbergen, “The effectiveness of query-specific hierarchic clustering in information retrieval,” *Inf. Process. Manage.*, vol. 38, no. 4, pp. 559–582, 2002.
- [10] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Upper Saddle River, NJ, 1988.
- [11] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, “A hierarchical monothetic document clustering algorithm for summarization and browsing search results,” in *Proc. Int. Conf. on World Wide Web*, 2004, pp. 658–665, ACM.
- [12] M.A. Hearst and J.O. Pedersen, “Reexamining the cluster hypothesis: Scatter/gather on retrieval results,” in *Proc. SIGIR-96*, 1996, pp. 76–84, ACM.
- [13] C. Carpineto and G. Romano, “Exploiting the potential of concept lattices for information retrieval with CREDO,” *J. UCS*, vol. 10, no. 8, pp. 985–1013, 2004.
- [14] H. Schutze and C. Silverstein, “Projections for efficient document clustering,” in *Proc. SIGIR-97*, 1997, Classification Methods, pp. 74–81.