BRIEF COMMUNICATION

Are Two Document Clusters Better Than One? The Cluster Performance Question for Information Retrieval

Robert M. Losee and Lewis Church, Jr.
Manning Hall, CB#3360, University of North Carolina-Chapel Hill, Chapel Hill, NC 27599-3360.
E-mail: {losee;churchl}@ils.unc.edu

When do information retrieval systems using two document clusters provide better retrieval performance than systems using no clustering? We answer this question for one set of assumptions and suggest how this may be studied with other assumptions. The "Cluster Hypothesis" asks an empirical question about the relationships between documents and user-supplied relevance judgments, while the "Cluster Performance Question" proposed here focuses on the *when* and *why* of information retrieval or digital library performance for clustered and unclustered text databases. This may be generalized to study the relative performance of *m* versus *n* clusters.

Introduction

To cluster or not to cluster; that is the question. Whether it is best to retrieve documents from one of two clusters, or evenly from the two clusters, or from an unclustered database, is discussed in the literature, but is not answered conclusively.

Clusters are groups of individual items that have been placed together or are treated as a unit, usually because of the similarity between the items that are placed into a cluster. A clustering algorithm assigns each individual item to a particular cluster. We will assume here that each entity is eventually located in exactly one cluster. A variety of clustering algorithms exist (Aldenderfer & Blashfield, 1984; Jain, Murty, & Flynn, 1999; Losee, 1990) and are illustrated in these sources.

Libraries have long brought books on similar topics together, this being a consideration in the development of many classification systems, such as the various Dewey decimal systems that are used throughout the world (Losee, 1993; Foskett, 1996). By placing similar items near each other in a library, browsing is improved, but not made perfect. Placing items in a single category often improves retrieval performance, but if one has interdisciplinary interests that run across traditional foci of library clustering, such as literature, history, social sciences, and philosophy, one may find traditional taxonomies and clustering to be something of a hindrance.

Clustering documents in automated retrieval systems has the potential to place similar documents near each other based on automatic determination of documents' features and machine-computed similarities. Underlying the effectiveness of this approach is whether the *clustering hypothesis* is true. Suggesting that "the associations between documents convey information about the relevance of documents to requests" (Jardine & van Rijsbergen, 1971), and later stated as "that relevant documents tend to be more similar to each other than to non-relevant ones, and therefore tend to appear in the same clusters" (Tombros & van Rijsbergen, 2001), the clustering hypothesis provides an empirically testable claim (Voorhees, 1985; Hearst & Pedersen, 1996; Leuski, 2001).

We believe here that clustering may help in some circumstances, but not in others. We choose not to address the clustering hypothesis directly, but instead ask the Cluster Performance Question:

> When does clustering of data improve information retrieval system performance, compared with performance obtained with no clustering?

We assume that there are circumstances where clustering may improve performance and other situations where clustering may decrease performance. This is related to the original cluster hypothesis, but has its focus on retrieval performance, rather than on an empirical question about relationships between documents and human needs: a question whose answer is certainly valuable but does not as directly address retrieval performance.

Received October 29, 2003; revised January 15, 2004; accepted January 15, 2004

 DOI: 10.1002/asi.20068

To simplify our discussion, we assume that our documents have a single binary term or feature that is employed by the user's query. Studying single-term systems permits the examination of the relationships between retrieval variables, relevant and nonrelevant documents, cluster characteristics, etc., rather than bringing into focus the relationships between document features, which is certainly of interest but obfuscates the clustering picture (Losee, 1998).

We also assume that query features have a positive relevance discrimination value. If this is not originally the case, we assume that the features are reparameterized so that features are positive discriminators.

Information Retrieval Performance

Most information retrieval systems provide either an ordered list of documents, which the user can examine until they decide to stop retrieving documents (Kraft & Waller, 1981), or a set of documents to be examined, with another set of documents being excluded from presentation to the user (Losee, 1998; Salton & McGill, 1983).

The precision of an ordered set of documents may be computed as the percent of relevant documents in the set. Precision may be computed for the increasingly large sets of documents if one wishes to show the relationship between precision and recall, the latter being the percent of relevant documents that are in the ordered set in question.

Precision may be computed for a fixed numer of documents. Many search engines, for example, provide retrieved sets of documents in groups of 10 documents per "screen." Studying precision-recall performance curves for a large search engine with access to billions of documents may be practically impossible. However, for "high-precision" searches, where the user wishes to see a few good documents, the precision for the first 10 documents, denoted as P_{10}, may be a useful measure of performance.

The precision for the first b documents retrieved may be computed as

$$P_b = P_{1,b} = \frac{1}{b}\left[\frac{r_1}{n_1}\min(n_1, b) + \frac{R - r_1}{N - n_1}(b - \min(n_1, b))\right]. \quad (1)$$

Here, R represents the number of relevant documents in the database of N documents, with r_1 relevant documents having the binary feature frequency of 1 from among the total of n_1 documents having the feature value of 1. The precision P_b is derived by computing the expected proportion of documents that are relevant from among the first b documents retrieved. We begin with retrieving documents with a frequency of 1. If there are at least b documents with a frequency of 1, then the expected proportion of relevant documents may be computed directly from this set. If there are documents without a 1 in the first b documents, we then examine the expected number of relevant documents from among those with a 0 as well as those with a 1 that would occur within the first b documents.

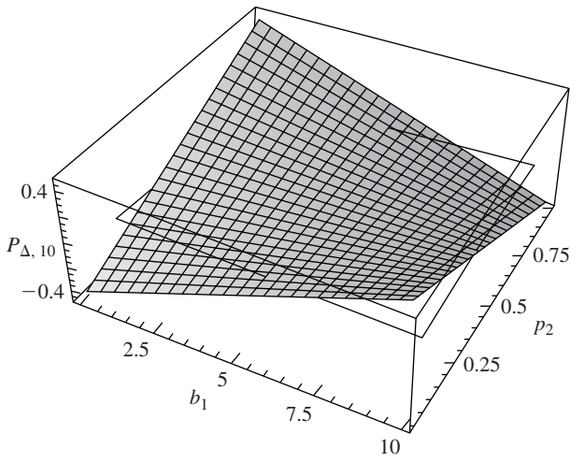

FIG. 1. Precision $P_{\Delta,10}$ for varying b_1, where $b_2 = b - b_1$, and for varying p_2, where $p_1 = 1 - p_2$. The break-even plane is shown halfway up the $P_{\Delta,10}$ axis. Performance of two clusters is much better than performance without clustering when b is maximized for a cluster and p is maximized for the same cluster.

The precision for the b documents in the merged set of documents from two clusters is,

$$P_{2,b} = P_{2,b_1,b_2} = \sum_{i=1}^{2}\frac{b_i}{b_1 + b_2}\left[\frac{r_{i,1}}{n_{i,1}}\min(n_{i,1}, b_i) + \frac{R_i - r_{i,1}}{N_i - n_{i,1}}(b_i - \min(n_{i,1}, b_i))\right]. \quad (2)$$

Extending the notation from the previous equation, R_i represents the number of relevant documents in cluster i, N_i represent the number of documents in cluster i, $r_{i,1}$ represents the number of relevant documents in cluster i with the binary feature frequency of 1, $n_{i,1}$ represents the number of documents in cluster i with the feature frequency of 1, and b_i represents the number of documents to be retrieved from cluster i.

We can address the question posed at the beginning concerning the utility of clustering by subtracting the values obtained from Equation (1) from the value of Equation (2), holding corresponding variables the same. In Figure 1, we compute this difference as $P_{\Delta,b} = P_{2,b} - P_{1,b}$. A positive value for $P_{\Delta,b}$ indicates that clustering outperforms nonclustering, with negative values indicating how much better nonclustering is compared with clustering.

The answer to the Cluster Performance Question is thus that clustering outperforms nonclustering when $P_{\Delta,b}$ is positive, given the constraints imposed above.

Results of Different Cluster Characteristics

We find that the distribution of documents between clusters can determine the relative performance of systems using clustering. If one cluster is empty, for example, precision for

the first b documents from the merged set is equivalent to the precision from an unclustered system. In Figure 1, we show the difference in precision $P_{\Delta,10}$ between two clusters and no clustering (the database is a single cluster).

In Figure 1 we arbitrarily assume that $R_1 = 7$, $R_2 = 9$, $N_1 = 20$, $N_2 = 25$, $n_{1,1} = 3$, and $n_{2,1} = 4$, and $r_{i,1}$ is computed for each cluster i to be consistent with the corresponding p values. We vary p_2 in the Figure, and set p_1 so that it is always equal to $1 - p_2$. Similarly, $b = 10$ and b_1 is varied in the Figure, and b_2 is always set to being $b - b_1$.

Figure 1 shows that as b_1, the number of documents extracted from the first cluster, increases (and b_2 correspondingly decreases), the relative advantage of two clusters over a single cluster increases for low p_2 (high p_1). For high values of p_2 (and correspondingly low values of p_1), we find that $P_{\Delta,10}$ increases as the number of documents from the first cluster decreases (and the number of documents from the second cluster increases).

This suggests that systems using two clusters will outperform systems using no clustering when most of the documents to be retrieved are in one of the clusters and that cluster has a high p value. This difference is maximized when, for cluster i, p_i approaches 1 and b_i approaches b.

Discussion

We have suggested that, in some circumstances, retrieving from clusters provides better performance than does retrieval without clustering. However, in other circumstances, the reverse may be true, and clustering may result in performance inferior to that obtained without clustering. Future retrieval system designers will need to address whether circumstances exist where clustering improves performance sufficiently, and equally importantly, whether the system they are proposing will use clusters in such a way that they increase query processing speed and throughput. While studying the clustering hypothesis leads one to directly study the effects of clustering on performance or the distribution of relevant documents across clusters, one can also study the performance of clustering by examining the p and b values of clusters, possibly using analytic tools such as Equations (1) and (2), and their difference, $P_{\Delta,b}$.

The measure of retrieval performance used in this study, $P_{i,b}$, was precision after retrieving b documents from i clusters, where b was assumed in Figure 1 to be 10. Other performance measures could be used for such a study, or the relationship between performance with this measure and performance with other measures might be studied (Losee, 2000). An earlier study on distributed information retrieval provides methods using Average Search Length as a performance measure that can address issues such as the clustering problems discussed above (Losee & Church, 2004). Using these techniques, we may generalize the study of the Clustering Performance Question to any number of clusters.

Future research in this area might help us understand the impact of several assumptions that were made above. Studying systems with more than one term may help us understand the interterm relationships and their affect on clustering, just as moving beyond two clusters to the more general case of n clusters will assist us in understanding more sophisticated relationships in the clustering world. However, the simplifying assumptions made above do provide us with significant observations and strong claims, and while they may appear to oversimplify the problem, they do provide us with powerful tools with which to address the core of complex cluster and retrieval problems.

References

Aldenderfer, M.S., & Blashfield, R.K. (1984). *Cluster analysis.* Beverly Hills, CA: Sage Publications.

Foskett, A.C. (1996). *The subject approach to information* (5th ed.). London: Library Association Pub.

Hearst, M.A., & Pedersen, J.O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In H.-P. Frei, D. Harman, P. Schauble, & R. Wilkinson (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 76–84). ACM. [AQ1]

Jain, A.K., Murty, M.N., & Flynn, P.J. (1999). Data clustering: A survey. *ACM Computing Surveys*, 31, 264–323.

Jardine, N., & van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7, 217–240.

Kraft, D.H., & Waller, W. (1981). A Bayesian approach to user stopping rules for information retrieval systems. *Information Processing and Management*, 17(6), 349–361.

Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. In *Proceedings of the Conference on Information and Knowledge Management (CIKM 01)* (pp. 33–40). ACM. [AQ1]

Losee, R.M. (1990). *The science of information: Measurement and applications.* San Diego, CA: Academic Press.

Losee, R.M. (1993). Seven fundamental questions for the science of library classification. *Knowledge Organization*, 20(2), 65–70.

Losee, R.M. (1998). *Text retrieval and filtering: Analytic models of performance.* Boston: Kluwer.

Losee, R.M. (2000). When information retrieval measures agree about the relative quality of document rankings. *Journal of the American Society for Information Science*, 51(9), 834–840.

Losee, R.M., & Church, L. (2004). Information retrieval with distributed databases: Analytic models of performance. *IEEE Transactions on Parallel and Distributed Systems*, 15(1), 18–27.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Tombros, A., & van Rijsbergen, C.J. (2001). Query-sensitive similarity measures for the calculation of interdocument relationships. In *Proceedings of the Conference on Information and Knowledge Management (CIKM 01)* (pp. 17–24). ACM. [AQ1]

Voorhees, E.M. (1985). The cluster hypothesis revisited. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 188–196). Association for Computing Machinery.

AQ1: Author: Location and full name of publisher.