A New Measure of the Cluster Hypothesis

Mark D. Smucker[1] and James Allan[2]

[1] Department of Management Sciences
University of Waterloo
[2] Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst

Abstract. We have found that the nearest neighbor (NN) test is an insufficient measure of the cluster hypothesis. The NN test is a local measure of the cluster hypothesis. Designers of new document-to-document similarity measures may incorrectly report effective clustering of relevant documents if they use the NN test alone. Utilizing a measure from network analysis, we present a new, global measure of the cluster hypothesis: normalized mean reciprocal distance. When used together with a local measure, such as the NN test, this new global measure allows researchers to better measure the cluster hypothesis.

Keywords: Cluster hypothesis, nearest neighbor test, relevant document networks, normalized mean reciprocal distance.

1 Introduction

Central to much of information retrieval (IR) is van Rijsbergen's cluster hypothesis: "closely associated documents tend to be relevant to the same requests" [1]. Early measurements of the cluster hypothesis pointed to the potential utility of cluster retrieval [2] and provided explanations of differing IR performance on different document collections [3].

Tombros and van Rijsbergen [4] recast the cluster hypothesis as not solely a property of a document collection but as a concern of a document-to-document similarity measure as applied to a document collection. With this view, we as designers of document-to-document similarity measures want to create similarity measures that *make the cluster hypothesis true* given a document collection and set of search topics.

In Tombros and van Rijsbergen's work, they created query-sensitive similarity measures (QSSMs). These similarity measures aim to focus similarity on the search user's topic. As such, what is considered similar to a given document changes for each search topic. Tombros and van Rijsbergen found that a QSSM has the ability to make the cluster hypothesis more true compared to similarity measures that ignore the user's query.

The current standard for measuring the ability of a similarity measure to make the cluster hypothesis true is Voorhees' nearest neighbor (NN) test [5]. The NN

L. Azzopardi et al. (Eds.): ICTIR 2009, LNCS 5766, pp. 281–288, 2009.

test measures the number of relevant documents found within rank 5 when a similarity measure ranks documents similar to a relevant document, which is effectively the same as measuring the precision at rank 5 (P5, the number of relevant documents found within rank 5 divided by 5).

Voorhees' NN test is notable for several reasons. The NN test says that what matters is whether or not non-relevant documents are ranked before relevant documents when documents are ranked for similarity to a given relevant document. Just because two relevant documents are very similar given a similarity measure does not preclude many non-relevant documents being more similar to the document. Perhaps most important though is that the NN test is comparable across different similarity measures, search topics, and document collections.

The NN test only requires relevant documents to *locally* cluster and cannot distinguish between a set of relevant documents that only locally cluster and a set of relevant documents that are also *globally* clustered. As such, the NN test may falsely report good clustering performance for query-biased[1] similarity measures. To see how this mistake is possible, assume we have a query that has many ($\gg 5$) relevant documents and a P5 of 1. If we query-bias the similarity until the query dominates over the given relevant document, then the rankings for every relevant document will be nearly identical and also have a P5 of 1. Using the NN test, we would declare the clustering performance to be excellent when in fact it could be very poor. The query may be high in precision but low in recall. Thus, all the relevant documents will be close to a few relevant documents but far away from the majority of relevant documents.

For some similarity measures and document collections, the NN test may fail to detect when relevant documents do cluster well. Wilbur and Coffee [6] found that the cluster hypothesis holds for the CISI collection in contrast to the NN test's negative conclusion [5]. Similar to Wilbur and Coffee's work, we utilized an earlier version of our methodology to measure the navigability of the find-similar interaction mechanism [7].

In this paper, we show that the NN test is an insufficient measure of the cluster hypothesis for a set of query-biased similarity measures. While the NN test works well as a measure of local clustering, it fails as a measure of global clustering. We present a new, global measure of the cluster hypothesis. We recommend that the use of the NN test be complemented with our test — each test tells us something different about how well relevant documents cluster.

2 A Global Measure of the Cluster Hypothesis

Our new measure of the cluster hypothesis is based on the shortest paths between relevant documents on a directed graph that we construct and call a *relevant document network*. We first describe the construction of the relevant document network and then we present our measure.

[1] We generically refer to similarity measures that bias similarity given the user's query as query-biased. QSSMs are one way to create query-biased similarity measures.

Relevant Document Networks. A document network is a graph where the nodes are documents and the edges of the graph represent some relationship between documents. For our purposes, we construct document networks as fully connected, weighted, directed graphs. An edge from a source document to a target document represents the similarity of the target document to the source when the source is used as a query. Documents and their similarities to each other have long been represented and manipulated in a graph theoretic fashion [1].

Rather than use the similarity measure directly as the edge weight, we set an edge's weight to be equal to the rank of the target document in the ranked results when the source document is used as a query. By weighting links in this manner, we gain the same benefits enjoyed by the NN test described above – notably the ability to directly compare across similarity measures, topics, and document collections. In addition, we exclude the source document from the ranked list of similar documents. We give documents not returned in a ranking an edge weight equal to the number of documents in the collection. Alternatively, one could give such edges a weight of infinity or equivalently not include the link in the graph. For a given source document, no two target documents have the same rank.

Rather than use the whole document network, we delete the non-relevant documents to produce a *relevant document network*. Figure 1 shows two examples of relevant document networks. Since each search topic has a different set of relevant documents, we construct relevant document networks on a per-topic basis.

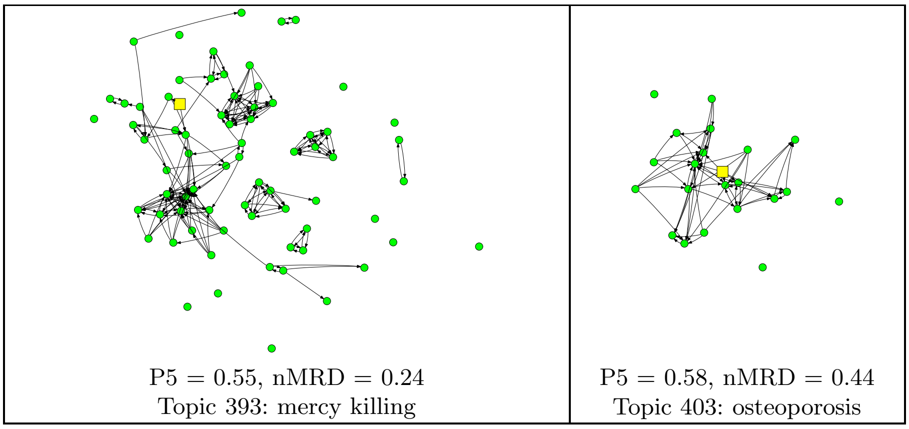


Fig. 1. Simplified depictions of relevant document networks for TREC topics 393 and 403. Each circular node is a relevant document. A link is drawn from a node to another node if the target node is within the top 5 most similar documents of the source node. The square node in each drawing represents a query likelihood retrieval using the topic's title as a query and is only shown for reference. The actual relevant document networks are fully connected, weighted, directed graphs. In this figure, the document-to-document similarity is "regular" with no weight given to the query/topic, i.e. $\lambda = 0$ (see Section 3).

Normalized Mean Reciprocal Distance. We propose as a new measure of the cluster hypothesis the *global efficiency* measure of Latora and Marchiori [8] applied to a given relevant document network. This measure is based on the shortest path distances between all pairs of vertices in the relevant document network.

This metric computes for each relevant document the normalized, mean reciprocal distance (nMRD) of all other relevant documents. The nMRD of relevant document R_i is calculated as:

$$nMRD(R_i) = \frac{1}{Z(|R| - 1)} \sum_{R_j \in R, j \neq i} \frac{1}{D(R_i, R_j)} \tag{1}$$

where R is the topic's set of relevant documents, $|R|$ is the number of relevant documents, $D(R_i, R_j)$ is the shortest path distance from R_i to R_j, and Z is the normalization factor. This metric varies from 0 to 1 with 1 being the best network possible. Because we allow no target documents to have the same rank, the best possible network for a given source document is a complete binary tree and thus:

$$Z = \frac{1}{|R| - 1} \sum_{i=1}^{|R|-1} \frac{1}{\lfloor \log_2 i \rfloor + 1} \tag{2}$$

For each topic, we average the nMRD over all the known relevant documents. Finally, for a test collection, we average over all topics to produce a final metric.

Looking again at the example relevant document networks in Figure 1, we see that precision at 5 (P5) reports that both topics 393 and 403 locally cluster relevant documents very well while the global clustering of the two topics is quite different. The normalized mean reciprocal distance (nMRD) reports that the relevant documents are globally much better clustered for topic 403 than for topic 393.

3 Document-to-Document Similarity Measures

In this paper, we use the well known language modeling approach to information retrieval to create a collection of document-to-document similarity measures. In our discussion, we refer to the document to which we are finding similar documents as the *source* document. For a given source document, we will call all other documents in the collection *target* documents.

In all cases, we build a query-biased, multinomial model, M_B, for a given source document and rank the remaining documents in the collection using the Kullback-Leibler divergence:

$$D_{KL}(M_B || M_D) = \sum_w P(w|M_B) \log \frac{P(w|M_B)}{P(w|M_D)} \tag{3}$$

where $0 \log 0 = 0$ and M_D is a smoothed, maximum likelihood estimated (MLE) multinomial model of the target document.

We generate a range of query-biased similarity measures by utilizing two ways to compute a model of the source document, M_S, and then by linearly combining this model with a MLE model of the given topic's query Q to produce a query-biased model M_B:

$$P(w|M_B) = \lambda P(w|Q) + (1 - \lambda)P(w|M_S) \tag{4}$$

where λ varies from 0 to 1 and controls the amount of query-biasing. While the query-biased similarity of Equation 4 is different than Tombros and van Rijsbergen's query sensitive similarity measure (QSSM) [4], which was a measure for vector space retrieval, the above formulation for language modeling retrieval captures the nature of QSSM.

We compute M_S by selecting differing amounts of the source document's text and then letting M_S be the MLE model of this selected text. At one extreme, we select all of the text in a document. When we select the whole document for M_S and set $\lambda = 0$, we produce what we call *regular* similarity, which is the most obvious form of document-to-document similarity and essentially treats a document as a very long query.

We also query-bias the similarity by how we select text from the document for M_S. Our second approach to query-biased similarity aims to capture the context of the query directly by only including the document text near query term occurrences. This "window" approach creates a MLE model of the source document text that consists of all words within a certain distance W of all query terms in the document. In effect, we place windows of size $2W+1$ centered over all query term occurrences in the document. For example, when $W = 5$, the window includes the 5 preceding words, the query term, and the 5 words following the query term. When selecting text in this fashion, if a document does not contain any query terms, the whole document is used.

Besides testing the "window" version of query-biased similarity alone by keeping $\lambda = 0$, we also take the query-biased model of the document that the "window" approach produces and mix this model with the MLE model of the query.

In summary, we have two ways of query-biasing the similarity. The first way mixes the query with a model of the source document M_S. The second way query-biases M_S by letting M_S be a MLE model of the text falling inside windows placed over query term occurrences in the source document. By comparing these two versions of query-biased similarity, we can see if the context captured by the windows holds an advantage over simply mixing the query with the whole document.

4 Experiments

We compared the NN test (P5) and the normalized mean reciprocal distance (nMRD) as measures of the cluster hypothesis on 150 TREC topics and 36 variations of document-to-document similarity.

To produce the 36 types of similarity, we took Equation 4 and investigated λ with values of 0, 0.1, 0.25, 0.5, 0.75, and 0.9. Besides utilizing the whole

P5

	Window Size W for M_S					
λ	All	15	10	5	2	1
0.90	0.45	0.45	0.45	0.45	0.45	0.45
0.75	0.46	0.47	0.47	0.47	0.48	0.48
0.50	0.51	0.52	0.52	0.53	0.52	0.50
0.25	0.56	**0.58**	0.57	0.57	0.54	0.50
0.10	0.51	0.53	0.53	0.53	0.50	0.47
0.00	0.32	0.40	0.42	0.44	0.44	0.42

nMRD

	Window Size W for M_S					
λ	All	15	10	5	2	1
0.90	0.15	0.15	0.16	0.16	0.16	0.16
0.75	0.17	0.17	0.17	0.17	0.18	0.18
0.50	0.20	0.21	0.21	0.22	0.22	0.21
0.25	0.27	0.29	0.29	0.29	0.28	0.25
0.10	0.30	**0.32**	**0.32**	**0.32**	0.30	0.27
0.00	0.20	0.27	0.28	0.29	0.29	0.26

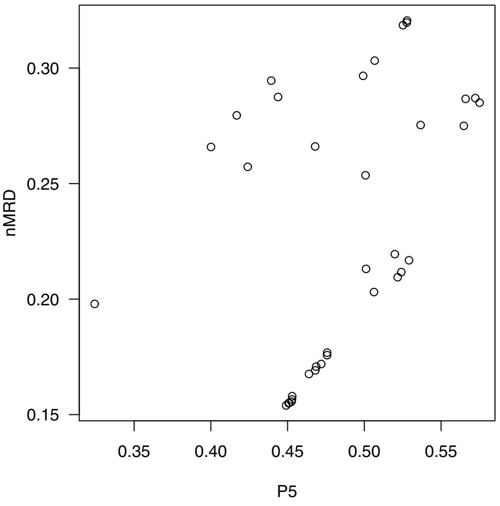


Fig. 2. Precision at 5 (P5) and normalized mean reciprocal distance (nMRD) measures of the cluster hypothesis for 36 variations of document-to-document similarity. The parameters λ and W refer to Equation 4 with "All" meaning the whole document is used to compute the source document model M_S. Scores without a statistically significant difference from the best scores are in **bold**. We measured statistical significance using the paired Student's t-test ($p < 0.05$). The plot on the right shows nMRD vs. P5.

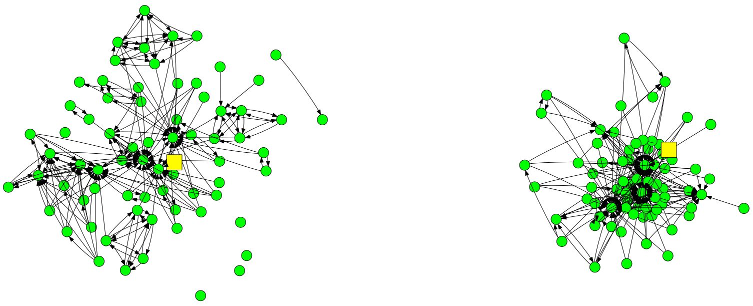
document to compute M_S in Equation 4, we also investigated window sizes W of 1, 2, 5, 10, and 15 words. The six settings of λ and six different window sizes for computing M_S produce the 36 similarity measures.

For our queries, we used the title field of TREC topics 301-450, which are the topics for TREC 6, 7, and 8. The document collection consists of TREC volumes 4 and 5 minus the Congressional Record. We smoothed the M_D of Equation 3 using Dirichlet prior smoothing with its parameter set to 1500. We truncated the query model M_B of Equation 4 to its 50 most probable terms. We stemmed using the Krovetz stemmer and used an in-house list of 418 stop words. We used the Lemur toolkit for our experiments.

5 Results and Discussion

Figure 2 shows the results for our experiments. While not shown here, we have found little difference relative to nMRD between P5 and the use of P10, P20, and average precision as local measures [9].

While generally a higher score for the NN test (P5) implies a higher score for the global measure (nMRD), there are numerous runs where the document-to-document similarity measure produced results with high P5 but with low nMRD. For example, the P5 measure has a value of 0.53 for the runs with a



$\lambda = 0.1$, P5 = 0.63, nMRD = 0.33 $\lambda = 0.25$, P5 = 0.57, nMRD = 0.19

Fig. 3. Simplified depictions of relevant document networks for topic 393 with $\lambda = \{0.1, 0.25\}$ (see Equation 4). Figure 1, shows topic 393 with $\lambda = 0$. Both Figure 1 and this figure compute M_S by using all the text of the source document.

window size $W = 5$ words and the λ values of 0.1 and 0.5, but when $\lambda = 0.1$, nMRD = 0.32 and when $\lambda = 0.5$, nMRD drops to 0.22. The NN test as a local measure of the cluster hypothesis is unable to measure the global clustering of relevant documents. If the NN test is used alone, it is possible to develop similarity measures that falsely appear to cluster documents well.

Nevertheless, to obtain a more complete view of the cluster hypothesis, both a global and local measure are needed. There are many similarity measures that produced relatively high nMRD scores between 0.25 and 0.3 while at the same time resulting in P5 scores ranging from 0.40 to 0.58. The nMRD measure is unable to detect local clustering of relevant documents.

Setting $\lambda = 0.1$ produced the best nMRD scores for all context sizes. Using a reduced context of 5, 10, or 15 words produced slightly better results than using the whole document (nMRD of 0.32 versus 0.30). For this document collection, it appears that there is some value to the window form of query-biased similarity although the majority of the benefit seems to come from giving the original query enough, but not too much weight.

The lower nMRD scores for the high values of λ are likely the result of a lack of diversity in the similarity lists across documents. Giving the query too much weight produces a ranking of similar documents that is more or less the same for all documents. From each document it becomes easy to traverse the relevant document network to a few relevant documents, but once at these documents, there is no easy way to travel to other relevant documents.

For topics such as topic 393 (Figures 1 & 3), we see that with a query-biased similarity, many of the outlying documents now have 2 or 3 of the query-similar documents as top ranked similar documents. These same outlying documents though have failed to gain connections to each other. Here it seems that query-biased similarity may be making the cluster hypothesis more true only by moving a few relevant documents closer to all relevant documents but not by helping all of the relevant documents get closer to each other.

While query-biased similarity has made the cluster hypothesis more true, the resulting connections between relevant documents are likely not robust to deletion of key, query-similar documents. If the query-similar documents did not exist in the collection, query-biased similarity might have had a less dramatic effect on the clustering of relevant documents. In addition to their global efficiency measure, Latora and Marchiori [8] have a local measure of efficiency that concerns itself with this question of robustness. In future work, we'd like to examine Latora and Marchiori's local efficiency measure and investigate to what extent similarity measures produce fault tolerant relevant document networks.

6 Conclusion

In this paper we presented a new measure the cluster hypothesis: normalized mean reciprocal distance (nMRD). This new measure is based on the shortest paths between documents on a relevant document network. In contrast to the NN test, which is a local measure of clustering, nMRD is a global measure of the cluster hypothesis. We examined 36 variations of document-to-document similarity and showed that the NN test is not a sufficient measure of the cluster hypothesis. Different similarity measures can score well on the NN test but have very different scores on the global measure, nMRD. To better determine the ability of similarity measures to make the cluster hypothesis true, both a global and local measure should be used.

Acknowledgments. This work was supported in part by the Center for Intelligent Information Retrieval, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors' and do not necessarily reflect those of the sponsors.

References

1. van Rijsbergen, C.J.: Information Retrieval, 2nd edn., Butterworths (1979)
2. Jardine, N., van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval. Information Storage and Retrieval 7(5), 217–240 (1971)
3. van Rijsbergen, C.J., Sparck Jones, K.: A test for the separation of relevant and non-relevant documents in experimental retrieval collections. Journal of Documentation 29, 251–257 (1973)
4. Tombros, A., van Rijsbergen, C.J.: Query-sensitive similarity measures for the calculation of interdocument relationships. In: CIKM 2001, pp. 17–24 (2001)
5. Voorhees, E.M.: The cluster hypothesis revisited. In: SIGIR 1985, pp. 188–196 (1985)
6. Wilbur, W.J., Coffee, L.: The effectiveness of document neighboring in search enhancement. IPM 30(2), 253–266 (1994)
7. Smucker, M.D., Allan, J.: Measuring the navigability of document networks. In: SIGIR 2007 Web Information-Seeking and Interaction Workshop (2007)
8. Latora, V., Marchiori, M.: Efficient behavior of small-world networks. Physical Review Letters 87(19) (October 2001)
9. Smucker, M.D.: Evaluation of Find-Similar with Simulation and Network Analysis. PhD thesis, University of Massachusetts Amherst (2008)