Measuring the Navigability of Document Networks

Mark D. Smucker and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
{smucker, allan}@cs.umass.edu

ABSTRACT

Browsing by similarity is a search tactic familiar to most people but one that the web unevenly supports. We are interested in user interface tools that augment the web with links to help users navigate from one relevant document to other relevant documents. We propose a combination of simple metrics to measure the navigability of document networks. These measures provide for low cost evaluation of the document networks formed by similarity measures and other link creation methods.

1. INTRODUCTION

After a long and tiring search, a user finally finds a web page relevant to the user's information need. While the page is relevant, it does not fully satisfy the user's need. How should the user proceed? If the page provides links to other pages, the user can follow those links. Alternatively, the user could follow links automatically produced by a tool that examines the page's content and provides links to similar pages. Tools that allow a user to request a list of documents similar to given document support the user interface feature we call *find-similar* [9].

Find-similar provides the search user a means to travel from one document to another. In effect, find-similar links together documents into a network, and just as a traveler in the physical world needs a good road system with direct routes, the search user needs find-similar to produce links that minimize the travel time to relevant documents. As applied to the web, find-similar aims to create a more navigable network by adding additional links to the existing document network that consists of web pages and hyperlinks.

A find-similar tool embodies some document-to-document similarity method. We would like to be able to test many variations of document-to-document similarity in a low cost manner. Testing different similarity measures with users is likely to be excessively expensive and likely to show little to no difference between similarity measures. Significant differences in retrieval quality can fail to be detected in user studies [4, 12].

The field of human computer interaction (HCI) has developed many methods of automated usability testing [5]. A premise of usability testing is that an interface exists to be tested. We would like measures of document-to-document similarity quality that are largely independent of the user interface otherwise we would need to test the cross product of interfaces and similarity measures.

Furnas [1] has developed a theory of *effective view navigation* that is related to our goal of efficient navigation from relevant document to relevant document. Furnas details his theory in terms of two types of graphs: a logical graph and a view graph.[1] The logical graph represents how objects, such as documents, are truly connected to each other. Furnas gives the web with its hyperlinks as an example of a logical graph. The view graph adds directed links to each node in the logical graph and represents the ways a user who is viewing the current node can immediately get to other nodes in the view. With find-similar, we are looking at ways to augment the logical graph and create a view graph that makes it easier for a user to find relevant documents.

To achieve effective view navigation, a system needs to be both efficiently view traversable (EVT) and view navigable (VN).

To be efficiently view traversable, Furnas requires two things. The first, EVT1, is that the views should be small, in other words, the out-degree of each node should be low when considering the view graph. The second, EVT2, is that the distance from each node to each other node on the viewing graph be short compared to the size of the overall structure.

Furnas' view navigability concerns itself with the "signage" aspects a of system. Links in the network need to provide good "residue" of the objects reachable via the link. Furnas' residue is similar to Pirolli's information scent [8]. In other words, the user needs the link labeled in a manner that provides a form of lookahead. At the same time, the label must be small. Simply providing a listing of everything reachable via the link would provide good residue but would result in too large of a label.

We see Furnas' use of out-degree as an approximation of the user's cost to use the link. As such, while the links in Furnas' graphs are unweighted, we weight each link in the network proportional to the time it takes a user to discover, evaluate, and travel a link.

One of our two measures of document navigability is based on the shortest paths between relevant documents. With regard to EVT2 (shortest paths), the question for information retrieval is not how easy is it to get from one document to

[1]We will use the terms network and graph interchangeably. In each case, we are referring to directed graphs, which consist of nodes and directed edges. Each directed edge connects a source node to a target node [3].

SIGIR'07 Web Information-Seeking and Interaction Workshop, July 27, 2007, Amsterdam, The Netherlands.

another, but how easy is it to get from a relevant document to other relevant documents. The searcher cares about the time to find relevant documents and not the time to travel between arbitrary documents. With a weighted document network, shortest paths now represent the optimal path for a user to follow between two documents.

A network with paths shorter than another network may actually be less navigable. For example, a randomly constructed network of low degree can have shorts paths between most nodes in the network. No user would be expected to navigate well in a random network.

Our other measure of network navigability aims to capture the quality of the similarity measure given the neighborhood it creates for a node. Hierarchical navigation networks such as the Yahoo! or DMOZ directories of web sites are examples of the difficultly of providing good node residue to achieve Furnas' view navigability for large document collections. The links at the top of these hierarchies are broad descriptions of the content available and offer little help in selecting the correct links. While we agree with the need for good link labels, with respect to the network structure, the network should be locally navigable. We are interested in document networks linked primarily at a local level — document to document. A good similarity measure produces links from relevant documents to other relevant documents. A random network would do poorly on this measure of navigability.

We propose using these two measures in combination to evaluate the navigability of a document network. When comparing two similarity methods, the better method should produce a network that is more navigable given both measures. We next discuss the two measures in detail.

2. PROPOSED MEASURES

Given a user's information need or search topic, a perfect similarity method for find-similar makes the topic's relevant documents most similar to each other. This is a restatement of the cluster hypothesis[6]. If a user finds a relevant document, and we have a "cluster hypothesis made true" similarity method, all a user needs to do is to request similar documents and the user will retrieve all of the relevant documents.

To measure the cluster hypothesis, Jardine and van Rijsbergen plotted the distributions of relevant pairs (R-R) and relevant and non-relevant pairs (R-NR) to visually determine the extent to which the cluster hypothesis was true [6]. This same procedure was examined in more detail by van Rijsbergen and Sparck Jones [13]. Griffiths, Luckhurst, and Willet replaced the visual inspection of the distributions with a measure of separation of the two distributions called the *overlap coefficient* [2].

Voorhees [14] pointed out that the relative frequency of very similar R-NR pairs is reduced by the large number of R-NR pairs in comparison to the number of R-R pairs. As an alternative, Voorhees proposed the *nearest neighbor* test, which counted the number of relevant documents found in the n nearest neighbors of a relevant document. Voorhees set $n = 5$. Voorhees' test is equivalent to examining the precision at 5 for the ranked lists produced by using relevant documents as queries. In place of precision at 5, any other retrieval metric such as average precision could be used in a similar manner. Using average precision would result in the computation of a mean average precision (MAP) for each given topic where each relevant document for that topic acts as a query. Voorhees' methodology has an added benefit that it is a measure that is more closely mapped to user notions of distance and separation.

We use Voorhees' methodology to measure the local quality of the document network. For each relevant document, we measure the average precision given the ranking of the document's neighbors formed by taking the weighted links as each neighbor's retrieval score.

A potential problem with the above mentioned measures of the cluster hypothesis is that they fail to accommodate the triangle inequalities that make the cluster hypothesis so appealing. We want to reward a similarity measure for making it easy to get from relevant document A to relevant document C by going first from A to relevant document B and then from B to C even if the similarity measure considers A and C to be dissimilar. To capture this feature of similarity and the value of navigating from document to document, we turn to a measure of the distance between documents measured on the network.

2.1 Document Networks

In a document network, the nodes represent documents in the collection and the edges represent a user's ability to traverse from a given document to another document via some user interface.

We aim to weight the links between documents in a manner that approximates the user's cost to find that link. Given a document-to-document similarity measure embodied by an implementation of find-similar or other user interface feature, for each document in a collection, we can compute a ranking of all other documents in that collection. While at best a crude approximation of user cost, we follow traditional information retrieval metrics and set a link's weight equal to its rank.

In some cases, we will have a document network but will not know the similarity measure. An example of this is the web graph. The links on a web page can be taken to be a ranking of the other web pages. For the links in the page, the top most link is given a rank of 1 and then the next link a rank of 2 and so forth. For many web pages, it may not be obvious from the HTML or even the visual layout of links what that proper ordering of links should be. Thus, an alternative that we follow in our experiments is to give all links a weight equal to the number of links plus 1 divided by 2, i.e. the average ranking. For example, each link on a page with 9 links will get a weight of 5.

Using document rank as our distance also provides us with another benefit. If we assume that shortest paths between relevant documents avoid passing through non-relevant documents, then we can delete the non-relevant documents from the graph and obtain the same results for the shortest paths between all pairs of relevant documents. Deleting the non-relevant documents produces what we term a *relevant document network*.

We obtain a substantial computational savings by deleting the non-relevant documents to form a relevant document network. For the relevant document network, we only need to calculate similarity information for the relevant documents rather than for all documents.

If non-relevant documents were to be on the shortest paths to relevant documents, relevant documents should have non-relevant documents as common neighbors. The cluster hy-

	Non-Relevant		
	10	20	100
Minimum	0.000	0.000	0.003
1st Quartile	0.018	0.024	0.039
Median	0.036	0.044	0.069
Mean	0.057	0.066	0.091
3rd Quartile	0.066	0.080	0.119
Maximum	0.593	0.717	0.543

Table 1: The average overlap coefficient among the top $N = 10, 20, 100$ ranked non-relevant documents in the nearest neighbors of relevant documents for TREC topics 301-450. For example, the mean fraction of non-relevant documents in common is 0.066 or 6.6% for the top 20 highest ranked non-relevant documents.

pothesis says that relevant documents share something in common to make them more similar to each other. In contrast, there is a limitless set of reasons that a document is non-relevant.

As a quick test of the extent to which non-relevant documents are common neighbors of relevant documents, we took the TREC topics 301-450 and we measured the overlap of the first N non-relevant documents occurring in the ranked lists produced by using a relevant document as a query. The document collection for topics 301-450 is comprised of newswire and government documents. While not a web collection, we feel it gives insight to this issue.

Our measure of overlap was the overlap coefficient:

$$\frac{|A \cap B|}{\min(|A|, |B|)}$$

where A is the set of N highest ranked non-relevant documents for relevant document A and similarly for document B. For each topic we computed the average overlap over all pairs of non-relevant documents and then computed summary statistics over all 150 topics. Table 1 shows that the amount of overlap is quite small with the mean overlap for $N = 20$ being 0.066 or 6.6% and three quarters of the topics have an overlap of 8% or less. Thus it appears that non-relevant documents play a role more akin to "noise" than as potentially useful stepping stones between relevant documents.

The assumption that a user will not navigate through non-relevant documents does not hold for document networks such as the web. On the web, links have a mixture of types. Some links go directly to other content rich pages while other links may go to a navigational page. Many navigation pages are not likely to be considered relevant pages in and of themselves. Imagine for example a web site that provides a find-similar link from each content page. The find-similar page is for navigational purposes and may link to a relevant page, but is not in itself a relevant page. By requiring paths to only go through relevant pages, for a similarity measure such as the web graph, we could cut off valid paths.

The relevant document network should only be used in situations where the document network is formed using a feedback-like technique such as find-similar. The relevant document network provides a reasonable upper bound on the shortest path where there is little sense in a user searching for relevant documents starting from a non-relevant document. While a non-relevant document may bridge two relevant documents, how would a user know how to decide between the good non-relevant documents and the bad ones? In a feedback situation, the user would be forced to "lie" to the system and judge a non-relevant document relevant.

2.2 Proposed Shortest Paths Measure

Given a weighted document network, we can efficiently compute shortest paths using Dijkstra's shortest paths algorithm or the Floyd-Warshall all pairs shortest paths (APSP) algorithm.

Distance on our weighted document networks represents the number of documents a user would need to examine by reading link labels such as document titles and summaries before reaching the other document. Other weighting schemes could approximate the individual costs of discovering, evaluating, and traversing links more closely.

Our proposed metric computes on a per topic basis, for each relevant document the mean reciprocal distance of all other relevant documents. Thus, the mean reciprocal distance of relevant document R_i is calculated as:

$$MRD(R_i) = \frac{1}{|R| - 1} \sum_{R_j \in R, j \neq i} \frac{1}{S(R_i, R_j)} \quad (1)$$

where R is the topic's set of relevant documents, $|R|$ is the number of relevant documents, and $S(R_i, R_j)$ is the shortest path distance from R_i to R_j. For each topic, we average the MRD over all the known relevant documents, and finally we average over all topics to produce a final metric. Because our minimum distance is 1, this metric ranges from 1 for the best possible score to 0 for the worst.

This measure is essentially the same as Latora and Marchiori's global efficiency measure [7]. Latora and Marchiori normalize the measure by dividing by the maximum possible efficiency in situations where the maximum efficiency is not 1.

3. EXPERIMENTS

We applied these two measures of navigability to three document networks: the web graph as represented by the wt10g TREC web collection, the document network formed on the same collection using a simple content based document-to-document similarity, and the combination of these two networks.

Soboroff [11] has shown the wt10g collection to have structural characteristics similar to the web. We used the TREC 2001 web ad-hoc topics numbered 501-550. Each topic defines a set of relevant documents. We do not use the topics' titles or descriptions in any way.

We constructed the web graph using the wt10g out_links file. To compute the document-to-document content similarity, we created a maximum likelihood estimated model of each document. We truncated each model to consist of only the document's 50 most probable terms. Using this model, we measure the similarity of the other documents using the KL-divergence. We used Dirichlet prior smoothing and set its parameter to 1500. We stemmed using the Krovetz stemmer and used an in-house list of 418 stop words. We used the Lemur toolkit for our experiments.

The content similarity network is a relevant document network and as such it only has links from relevant documents

	Mean Average Precision		
	Web	Content Sim.	Mix
Minimum	0.000	0.003	0.003
1st Quartile	0.000	0.045	0.040
Median	0.000	0.073	0.067
Mean	0.002	0.101	0.093
3rd Quartile	0.002	0.140	0.131
Maximum	0.022	0.375	0.375

Table 2: The mean average precision for the three document networks where "Mix" is the combination of the web and content similarity networks.

	Mean Reciprocal Distance		
	Web	Content Sim.	Mix
Minimum	0.000	0.002	0.004
1st Quartile	0.003	0.022	0.029
Median	0.004	0.034	0.040
Mean	0.005	0.064	0.071
3rd Quartile	0.006	0.052	0.061
Maximum	0.024	0.750	0.750

Table 3: The mean reciprocal distance for the three document networks where "Mix" is the combination of the web and content similarity networks.

to other relevant documents as described in Section 2.1. We only included content similarity links that had a weight of 100 or less.

Tables 2 and 3 show the results. These tables show the summary statistics across the 50 topics for each measure. For example, in Table 2 the web has at least one topic for which the mean average precision (MAP) was 0.000. A topic with a MAP measure of 0.000 means that the average relevant document has no hyperlinks to any other relevant documents. For example, topic 548 has only two relevant pages. Neither page links to each other. Thus, for topic 548, each page has an average precision of 0 and the mean average precision for the topic is 0. This does not mean there isn't a path from relevant document to relevant document. Also, it may not be the case that the worst or best score for one network is the same topic that is the worst or best for another network.

The web alone does not appear to provide good navigability either locally or globally. The content similarity links appear to be much more navigable. This echoes our other findings where we found that adding 10 content similarity links to web pages brings relevant documents closer to each other [10]. In this other work we gave all links a weight of 1 and only looked at distance on the graph between relevant documents. While a small effect, compared to content similarity alone, combining the two networks hurts the local navigability (MAP) while helping the global navigability (MRD).

4. CONCLUSION

We have proposed measuring the navigability of a document network using two measures. The nodes in the network represent the documents in the collection and the directed links represent the ability of a user to traverse from a source document to a target document. The weight of a link is set proportional to the user's cost to find, evaluate, and traverse the link. One measure captures a local and the other a global quality of the network. The local quality of a network can be measured as follows. For each relevant document, we rank a document's neighbors by their link weights and measuring the average precision of this ranking. The measure of local quality is the mean average precision for the relevant documents. The global measure captures the cost to follow the shortest path, navigating from a relevant document to another relevant document. For each relevant document, we measure the mean reciprocal distance to all other relevant documents. The overall measure is the average of these mean reciprocal distances. Together, these two measures should give us a good understanding of the navigability of a document network and allow us to design similarity methods that construct more navigable networks.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, and in part by NSF Nano # DMI-0531171. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] G. W. Furnas. Effective view navigation. In *CHI '97*, pages 367–374. ACM Press, 1997.

[2] A. Griffiths, H. C. Luckhurst, and P. Willett. Using interdocument similarity information in document retrieval systems. pages 365–373, 1997.

[3] N. Hartsfield and G. Ringel. *Pearls in Graph Theory*. Academic Press, Inc., San Diego, 1990.

[4] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR '00*, pages 17–24. ACM Press, 2000.

[5] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv.*, 33(4):470–516, 2001.

[6] N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, 1971.

[7] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87(19):198701, Oct 2001.

[8] P. Pirolli. *Information Foraging Theory*. Oxford University Press, 2007.

[9] M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *SIGIR '06*, pages 461–468. ACM Press, 2006.

[10] M. D. Smucker and J. Allan. Using similarity links as shortcuts to relevant web pages. In *SIGIR '07*. ACM Press, 2007. Poster, to appear.

[11] I. Soboroff. Do TREC web collections look like the web? *SIGIR Forum*, 36(2):23–31, 2002.

[12] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR '01*, pages 225–231. ACM Press, 2001.

[13] C. J. van Rijsbergen and K. Sparck Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29:251–257, 1973.

[14] E. M. Voorhees. The cluster hypothesis revisited. In *SIGIR '85*, pages 188–196. ACM Press, 1985.