Using LSI to evaluate the quality of hypertext links *

James Blustein ⟨jamie@csd.uwo.ca⟩
Robert E. Webber ⟨webber@csd.uwo.ca⟩

Department of Computer Science, University of Western Ontario
London, Ontario, N6A 5B7

Useful hypertext is constrained by the need for users to be able to find documents about similar topics without extensive navigation. We show how examining the properties of a graph built by a document's hypertext links can be used to evaluate the usefulness of the document. To formally measure the quality of hypertext linking in a corpus, we compare the semantic similarity of pairs of documents with the minimum number of links between their corresponding nodes in an analogous hypertext graph. We use the measure of document-to-document similarity computed using latent semantic indexing as our measure of semantic similarity. Our method has been applied to a corpus composed of Usenet messages.

Introduction

We present a new method for evaluating the usefulness of hypertext links and we present experiments using this method. Each experiment made use of the same corpus of 1608 documents from approximately 320 authors. The documents were drawn from the main Usenet newsgroup about computer graphics. A typical document in the corpus had 200 words.

Hypertext is any form of 'non-sequential writing — text that branches and allows choices to the reader, best read at an interactive screen' [1]. Linked hypertext is the most prevalent form of hypertext today. In such a document, users navigate between chunks of text by following links. Our goal is to create linked hypertext that will be useful for browsing. We compare the semantic closeness of documents with the number of links in

*Presented at ACM SIGIR *IR and Automatic Construction of Hypermedia: a research workshop*, Maristella Agosti and James Allan, eds. July 1995.

the shortest path between them as a measure of how well the links create a structure for browsing.

What is good hypertext

Useful hypertext is constrained by the need for users to be able to find documents about similar topics without extensive navigation. Hypertext navigation can be difficult when users must choose from an overwhelming number of links or when they must follow many links to read related documents. Furthermore, practical hypertext documents require most text chunks to be reachable from each other.

If links make explicit connections between documents that already share the same topic, i.e., are *semantically close*, then the most closely related documents will have direct links between them. Similarly, if two documents are directly linked, then they should be semantically close. If a user must traverse many links to go from document A to document B, then A and B should be semantically further than another pair of documents that are fewer links away from each other.

Automatic conversion into hypertext

A major obstacle to more efficient use of information is the difficulty of converting existing documents into hypertext documents. Unlike our experiment, much of the work in this area requires the use of highly structured text [2, 3]. When people convert unstructured text into hypertext documents, they must form links between different chunks of that text. In order to make these links, they must read and understand the text well enough to make connections between the parts. Since it is infeasible to have a human create thousands of links and evaluate them all [4], we have explored methods for automatic link creation that do not require a person to read or understand the text to build a hypertext document.

Experimental design

By comparing the link measure and semantic closeness, we formally measure the quality of hypertext linking in a corpus (see Figure 1). If we think of the text chunks in a hypertext as nodes and the links as edges, then the link measure is reducible to the graph problem of computing the shortest path between two nodes. The *shortest path* from node A to node B contains the minimum number of edges that must be traversed to go from A to B. The *link measure* between documents A and B is exactly the shortest path length between the corresponding nodes in the hypertext graph.

We distinguish two major types of links: structural and semantic. *Structural links* act as cross-references already present in documents. They are quite straightforward and their quality was not evaluated in the current experiments. The connections formed by *semantic links* are based on the ideas incorporated into the documents and are the subject of the current experiments.

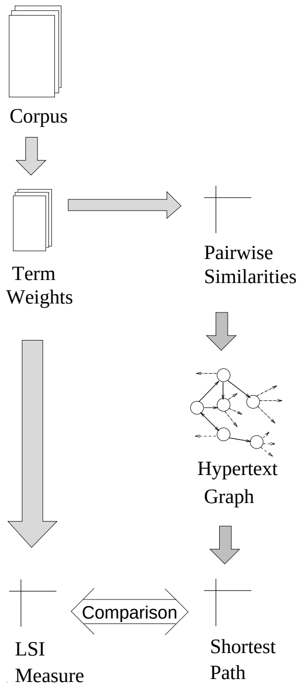

Figure 1: Experimental Design

In our experiments, we created a graph representation [5] of hypertext links by applying three global weighting schemes — inverse document frequency (Idf), entropy, and GfIdf [6] — in combination with local log term weighting. One edge from each document to exactly the n most similar documents (for a given n from 1 to 5) were created in the corresponding graph. We computed the similarity measure as the cosine of weight vectors built using each of the weighting schemes mentioned above. For example, in one of the experiments we computed the Idf and log weights for all terms in each document and connected each document to the three most similar other documents. The combinations of three weighting schemes and five outdegree values resulted in 15 different experiments.

We compared the semantic closeness of pairs of documents with the link measure between their corresponding nodes in the analogous hypertext graph. We also considered the average path length between nodes as a measure of utility. We use the measure of document-to-document similarity computed using latent semantic indexing (LSI) to determine the semantic closeness of pairs of documents [6].

The LSI document-to-document similarity measure is based on the document vocabulary as it is used in the corpus — not in any external source, e.g. a thesaurus. LSI was used on the basis of a report in the literature [6]. However, recent results [7] indicate that there are some real differences between the performance of LSI and other top IR systems. It is unclear how the results of the experiments may have been affected by the use of LSI.

Other IR systems could be used in place of LSI for the experiment.

Results

There are two considerations in selecting an appropriate conversion method: 1) the correlation between semantic closeness and link measure and 2) the degree of connectivity in the resulting hypertext. Table 1 shows the comparison of measures for the hypertext built using Idf and log weights.

The second column shows the coefficient of correlation [8] between the semantic closeness and link measure for all document pairs that had a path between them. This measure tells us whether closely related documents are closely linked. Because high document-to-document similarities lead to small path lengths, the correlations between the link measure and semantic closeness are less than zero.

Hypertext graphs with single outdegree typically have large correlations, but they connect only a tiny fraction of the documents in the corpus. A hypertext system geared towards searching and browsing must have paths between many documents. The final column presents the *coverage*, i.e., the fraction of document pairs that have paths between them.

n	Correlation	Coverage
1	-0.7323	0.17%
2	-0.2775	1.18%
3	-0.0216	47.51%
4	-0.0164	76.59%
5	-0.0200	87.31%

Table 1: Properties of hypertext created using Idf-Log weighting with n links from each document

Discussion

In a real hypertext system a coverage figure of less than 50% means that there is no way of navigating between most documents. This figure is a concrete way of describing the sparseness of the corresponding graph. The smallest graph in our experiments — 4470 document pairs — was the combination of the Idf and log term frequency weights with an outdegree of

1.

The differences between the LSI measure and link measure can lead to significant differences in correlation figures. LSI provides a measure of similarity for every document pair but hypertext graphs may contain document pairs with no path between them. Because the LSI similarity measure is a real number and the shortest path is a whole number, the link measure is at best an approximation of the semantic closeness of two documents. Small differences in how the hypertext graph is made can have significant consequences for a linearly based comparison of the measures, viz. correlation.

Clearly, there is more information about document-to-document similarity from the LSI measure than from the shortest path data. An examination of the structure of the graphs reveals that when more links are permitted path lengths tend to shrink as related documents become closely connected. In our corpus, the greatest path length shrinkage occurred between graphs with outdegree three and four. Shorter maximal paths will permit more browsing as suggested by Raymond and Tompa [9]. As part of our future investigation, we want to perform a more rigorous analysis that will allow us to evaluate differences caused by the weightings used and the discretization caused by the conversion to a graph.

There is interest in hypertext versions of Usenet articles although they are shorter than those often used in IR experiments. The length of our documents is not as critical as in many other IR experiments since we are comparing all documents to each other. The hypertext we create can be considered a single document composed of many chunks, where each chunk is a document in the corpus. Both the computation of semantic similarity and of weighting schemes used the cosine measure to normalize for document length.

Conclusion

We show how examining the properties of the graph built by a document's hypertext links can be used to evaluate the usefulness of that document. This evaluation is primarily based upon a comparison of semantic closeness and the link measure for all pairs of linked documents. We also examine the maximal shortest path and the sparseness of the graph. This approach is objective and particularly well-suited to large corpora. Our tests indicate that a combination of global Idf and local log weights produces a better linked hypertext than the other methods tested.

References

[1] Theodor Holm Nelson. *Literary Machines*. The Distributors, 90.1 edition, 1990. pp. 0/2 – 0/3.

[2] Eanass Fahmy and David T. Barnard. Adding hypertext links to an archive of documents. *The Canadian Journal of Information Science*, 15(3):25 – 41, September 1990.

[3] Richard Furuta, Catherine Plaisant, and Ben Shneiderman. A spectrum of automatic hypertext constructions. *Hypermedia*, 1(2):179 – 195, 1989.

[4] David Ellis, Jonathan Furner-Hines, and Peter Willett. On the creation of hypertext links in full-text documents: Measurement of inter-linker consistency. *The Journal of Documentation*, 50(2):67 – 98, June 1994.

[5] Jacques Savoy. A learning scheme for information retrieval in hypertext. *Information Processing & Management*, 30(4):515 – 533, 1994.

[6] Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229 – 236, 1991.

[7] Jean Tague-Sutcliffe and James Blustein. A statistical analysis of the TREC-3 data. In *Text Retrieval Conference*, Gaithersburg, MD, USA, November 1994. National Institute of Standards and Technology.

[8] Mark L. Berenson, David M. Levine, and Matthew Goldstein. *Intermediate Statistical Methods and Applications: A Computer Package Approach*. Prentice-Hall, 1983.

[9] Darrell R. Raymond and Frank Wm. Tompa. Hypertext and the Oxford English Dictionary. *Communications of the ACM*, 31(7):pp. 877 – 878, July 1988.