

# Improving Authoritative Sources in a Hyperlinked Environment via Similarity Weighting

Jonathan D. Herbach  
Princeton University, USA  
Email: jherbach@cs.princeton.edu

## ABSTRACT

Recent literature demonstrates that the network structure of a hyperlinked environment can be an effective source for inferring the importance of content in documents. To compensate for the problems of pure hyperlink connectivity analysis, we introduce consideration of document content by creating a variant of a known algorithm. Using Web data, we investigate how incorporating document similarity modifies how link structure determines the importance of documents.

**KEYWORDS:** authoritative sources, connectivity, ranking, similarity, search engines, importance, information retrieval

## INTRODUCTION

Our research focuses on improving the Hyperlink-Induced Topic Search (HITS) algorithm for determining importance and ranking a set of documents accordingly. [4] HITS determines importance of documents (e.g., pages on the Web) not by their content (e.g., frequency of query terms in document), but by the hyperlink structure of the collection. The underlying assumption is that when a document links to another, there is good reason for the hyperlink. Essentially, documents referenced most frequently are considered ‘better’ authorities and therefore more important.

Work by Bharat and Henzinger modifies HITS to address the *topic drift* problem: hyperlink structure can generalize a specific query. [1] If the topology of the local graph is well-connected and not relevant to a narrow query topic, HITS may determine that pages within the local graph are more important because HITS considers all hyperlinks equally important. Chakrabarti, et al. investigate a similar problem: preventing consideration of purely navigational links. [2]

## ALGORITHM

We modify the HITS algorithm, calling it HITS-SW because we assign similarity weights to each hyperlink and consider links to be of unequal importance. Each hyperlink’s importance is determined by the similarity between the content of

the source and destination documents, where weight assignments are real numbers between 0 (dissimilar) and 1 (similar). HITS-SW then considers a subgraph of a collection of pages based upon a desired range of similarity weights.

As illustrated in figure 1, dotted hyperlinks indicate those not matching the desired similarity range. “Q” pages contain the query terms. On the subgraph of desired edges, white pages are neighbors of query pages, whereas grey pages are not. Grey hyperlinks leave or enter grey pages. In this example, HITS-SW considers only white pages and bold hyperlinks; HITS would consider all (including grey and dotted) pages and links, as all hyperlinks are internal. HITS-SW performs the same calculation as HITS but on this different graph. [4]

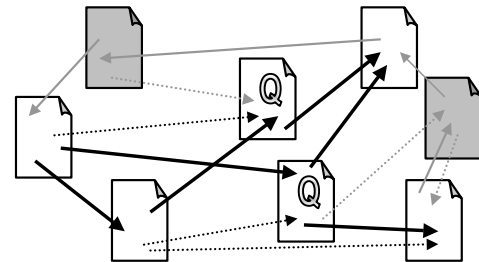


Figure 1: Web with Similarity Weights for HITS-SW

## METHOD

Our full study uses several classic methods to determine similarity between documents; here we report just one. The distinct words and word-pairs found in pages were stored in binary term vectors normalized by an arithmetic mean coefficient. [5] All 150,000 pages within the *princeton.edu* domain normally indexed by the local search engine were used as data for the evaluation of HITS-SW. The text from each page was retrieved, cleaned using a standard stopword list, and stemmed using an algorithm by Porter. [6]

We ranked 29 queries with HITS-SW using links weighted greater than (*hi*) and less than (*lo*) the median weight value. As a control, each query was also ranked with HITS-SW using randomly assigned similarity weights. The control tests were run 25 times for each query, with each simulating an arbitrary “similarity weighting function” by which 50% of the edges were declared “within the threshold” for HITS-SW. The goal was to test whether HITS-SW-like results could be

