CSE494 Project Report - Part B

J. Benton

March 30, 2004

1 Introduction

The objective of this project is to compare the resulting rankings given by Authority/Hubs (A/H) with the vector space search, PageRank + vector space similarity with Authority/Hubs and to comment on the effects of changing parameter values in both algorithms.

2 Design and GUI

A/H and PageRank both have two associated classes, a class that implements a single iteration of the calculations and another class that executes the iterations. The single iteration class checks for convergence of the involved rank vectors as a halt condition. The executing classes indicate the current status of the query. This status is pulled by the GUI and updated on a status line.

Figure 1 shows the GUI with the current status of search located at the bottom. Parameters can be changed based upon the type of search chosen by the radial buttons. A progress bar indicates how far the current status has progressed (the number of iterations is used as a metric). When no status is necessary, the progress bar is removed from view to allow more room for the search results on the screen. Results of the search are shown in the large text box below the top panel as shown in Figure 2.

3 Analysis of Authorities/Hubs

Vector space similarity ranking is calculated based upon the number and frequency of terms within a document. In contrast, the Authorities/Hubs (A/H) ranking method depends upon the number of links pointed to or pointed from a particular page. Authorities/Hubs should be calculated on a set of documents that are determined to be relevant to a particular query. To estimate relevancy, the vector space similarity calculation is performed. After this, the top K results (the "root set") and the pages pointing to and pointing from those results are combined. This provides a set of documents that can be calculated upon.

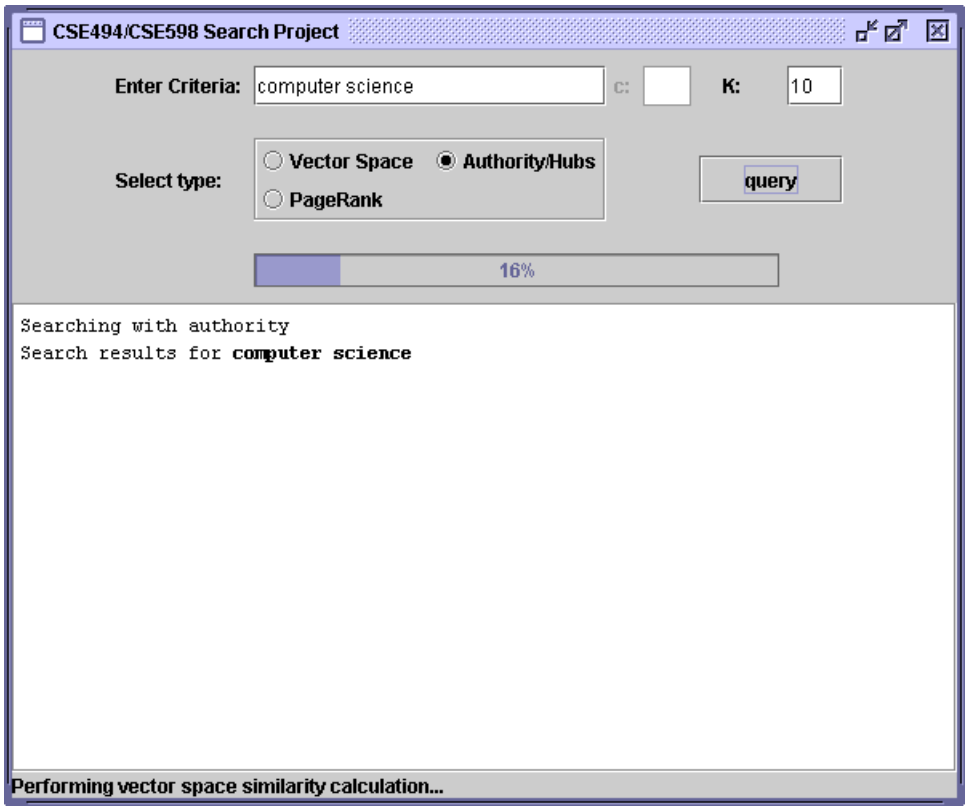

Figure 1: The GUI used for performing searches. Note the status bar indicates the current status of the search. Also, text boxes exist to change search parameters.

A/H tends to return more relevant documents than vector space similarity. But since the A/H ranking method depends upon the vector space similarity calculation, we provide an analysis of the vector space similarity method with A/H on two illustrative queries from the given crawl. The queries "src" and "fall semester" together show how important accurate relevancy estimation is when calculating the A/H rankings.

For the query "src" vector space similarity returns, out of the top 10 results, 8 relevant documents. In contrast, "fall semester" returns 0 relevant documents from the top 10 of the vector space query. Figures 3 and 4 show this along with the A/H results in terms of precision. For these two queries, there is a correlation between the rise and fall of precision in the A/H ranking and the vector space similarity ranking. Both figures also show when vector space is below optimum precision, the Hubs ranking tends to have a better precision than the Authorities ranking. However, when the vector space ranking gives a

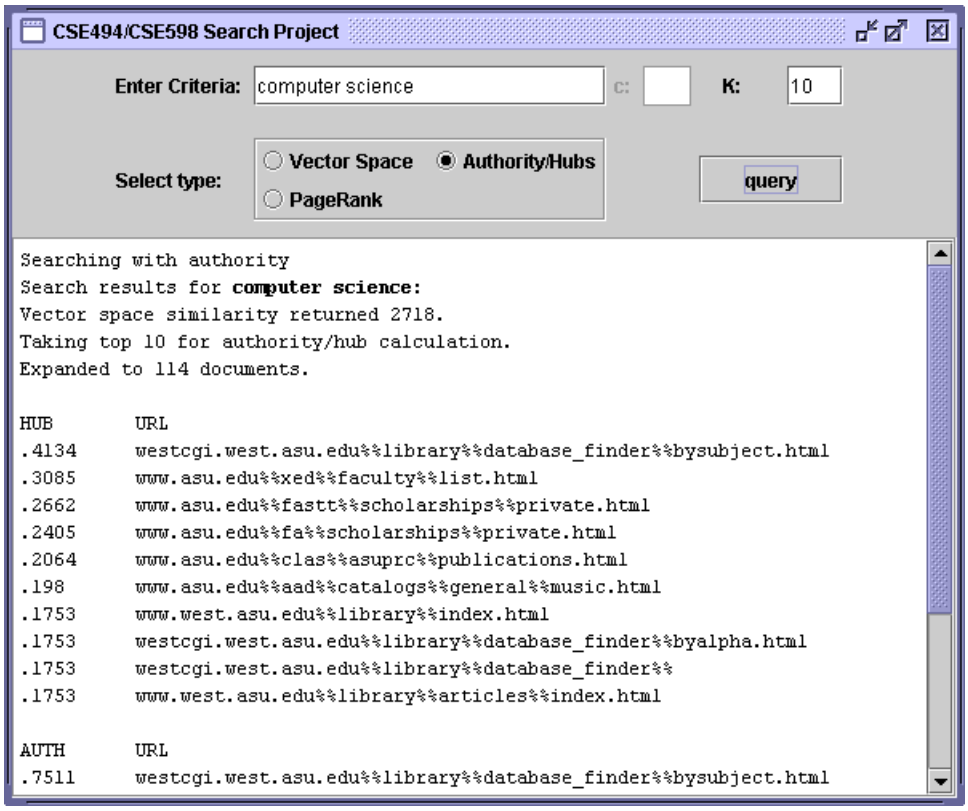

Figure 2: Search results are shown in the large text box below the top panel.

more accurate relevancy estimation, the Authorities ranking is better than or equal to the Hubs ranking in terms of precision.

Because Authorities/Hubs lacks the incorporation text search, it depends upon another method (in our case vector space similarity) to do this. The algorithm depends on the relevance of the documents as a user might.

4 Analysis PageRank + Vector Space

Like the Authorities/Hubs ranking algorithm, PageRank depends on another method for determining whether a particular query satisfies text relevancy (i.e. the query text is associated in some way with the returned documents). In our implementation, PageRank differs in that a final ranking is determined using a weighted combination of vector space similarity, that is $w \times PageRank + (1-w) \times Similarity$ where w is a weight between 0 and 1. We normalize the PageRank so that it fits between 0 and 1 by ensuring that the sum of the PageRanks for

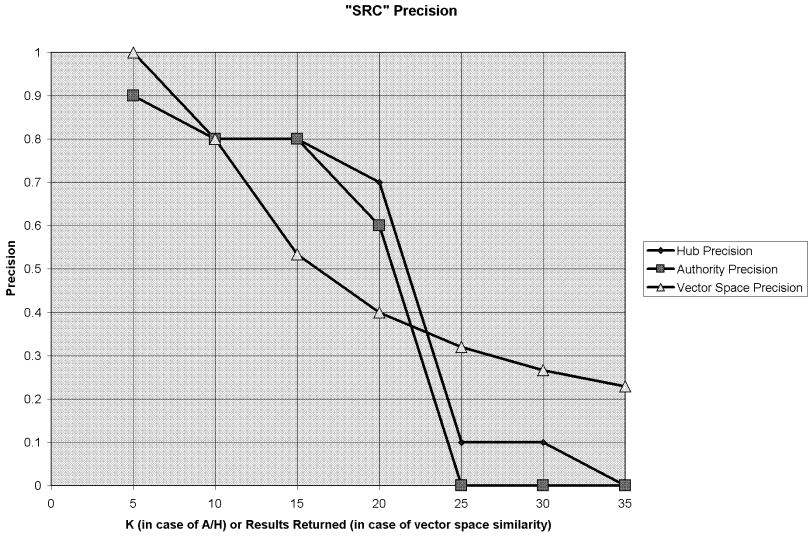

Figure 3: Comparison of vector space and A/H for the query "SRC".

the corpus adds to 1. Because of this, to allow PageRank to have a notable effect, weights must be high. Figure 5 shows the effect that varying the weight has on two different queries.

We can see that increasing the PageRank weight has a negative effect on the "parking decal" query. The query returns 1136 vector space results. The "networks" query returns 181 results. For the "parking decal" query, as PageRank increases in weight, more popular documents (such as the page "www.asu.edu") receive a greater ranking despite their irrelevancy. On the other hand, "networks" returns fewer vector space results. PageRank benefits this query since popular pages are not included in documents containing the term.

By decreasing the c value in the PageRank calculation, we are giving more credit to the random surfer K matrix when creating M^*. When we increase, c the distributed importance of a page is being increased. The effect of this is shown in Figure 6. Here we see that the random surfer matrix is important to include but should not be given a very high weight.

To compare A/H ranking with the PageRank + Vector Space ranking, we took the precision of four queries for each. Table 1 shows the results. We can see that for most queries, PageRank + Vector Space returned more relevant

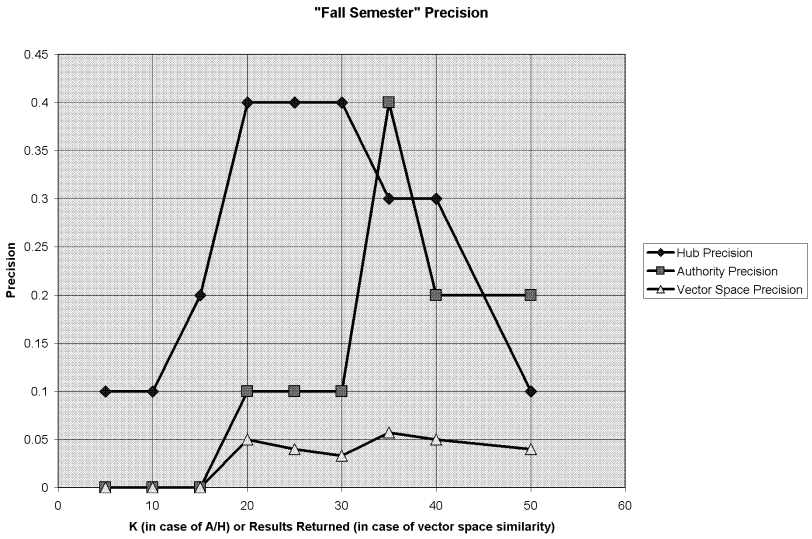

Figure 4: Comparison of vector space and A/H for the query "fall semester".

documents more often[1] Also, when A/H fails to return any relevant documents PageRank succeeds.

[1]Used weight of 0.998 for PageRank and $c = 0.9$. Used "root set" of 10 for A/H.

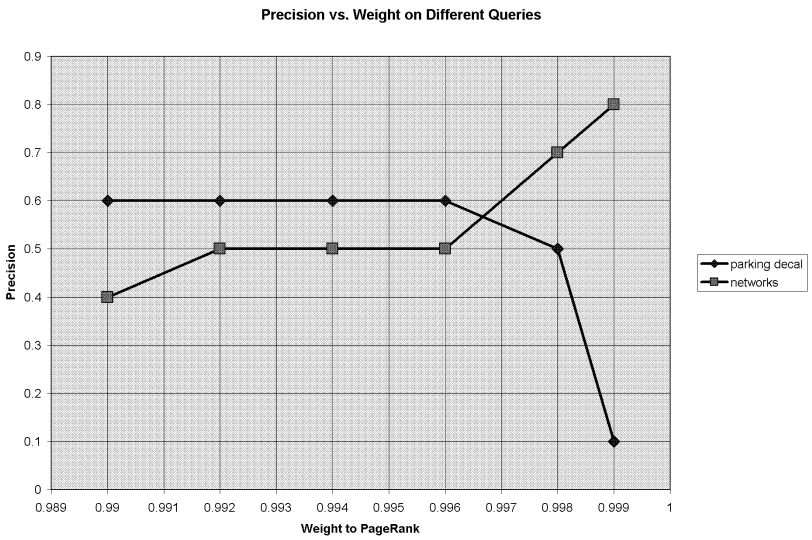

Figure 5: On different queries, the same weight given to PageRank has a different effect.

query	**A/H Hubs**	**A/H Authorities**	**PR + VS**
transcript	0	0	6
networks	0	0	6
fall semester	2	4	5
parking decal	7	9	7

Table 1: The precision for A/H and PageRank + Vector Space on various queries.

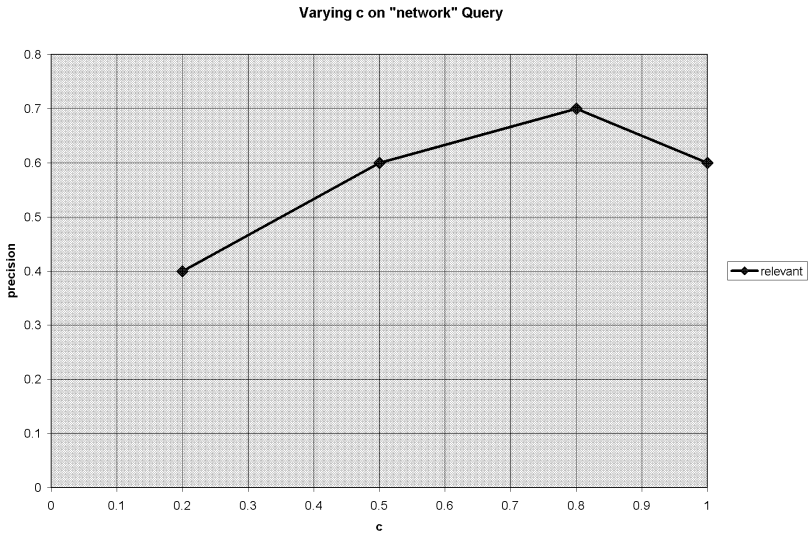

Figure 6: The random surfer matrix (K), when given a greater weight (i.e. when "c" is small), the precision of the results is poor. It is the case, however, that the matrix provides some precision when given some weight (i.e. when "c" is larger but not "too large").