# CSE 494 Project B – Task 3

**Garrett Wolf**

## Introduction

The main purpose of this project task was for us to implement the Authorities/Hubs ranking algorithm. Once implemented, we were asked to vary the size of the roots set and observe the results. We were provided with a link matrix from the crawled files and code allowing us to get the inbound and outbound links for each of these files. The link matrix was used to implement the Authorities/Hubs algorithm and integrate it into the application from Project A. This write-up will provide a background discussion on the Authorities/Hubs ranking algorithm, a comparison of the Authorities/Hubs results to those of the Vector Space model, an analysis of the affects on the results from changing the root size, and a judgment on whether the Authority or Hub results are more relevant.

## Authorities/Hubs Algorithm Background

The Authorities/Hubs ranking algorithm is based on the idea that a page that is linked to by a lot of important pages is important and is called an Authority. Similarly, a page that links to a lot of important Authority pages is also important and is called a Hub. This type of model allows importance and rank to be propagated. In the Authorities/Hubs model, a page's authority score is calculated as the sum of the hub scores from all the pages pointing to that page. A page's hub score is calculated as the sum of the authority scores from all the pages that the page points to. These scores can be computed iteratively until they begin to converge. In this implementation, the user's query is first processed using the Vector Space ranking model. Next, the top *k* results from the Vector Space computation are used to construct a root set of pages. These pages are then added to a base set and expanded by adding any page that is pointed to by a page in the root set. Any page that points to a page in the root set is also added to the base set although a maximum of 50 of these pages will be added to ensure the base set does not become too large. The base set is then used to create an adjacency matrix depicting the links between pages in the base set. For this piece, the provided link matrix and extraction methods were used to find a page's inbound and outbound links. Next, the authority and hub scores are initialized to 1 for each of the pages. Once initialized, the iterative loop begins by first making a copy of the authority scores. Then the new authority scores are calculated as the sum of the hub scores for all the pages pointing to a page. After calculating the new authority scores, each page's hub score is calculated as the sum of the authority scores for all the pages that the page points to. Once complete, both the authority and hub scores are normalized. Using the normalized authority scores and the original copy of the scores, a residual is calculated to determine if scores have converged. The residual is calculated using the formula below:

$$residual = \sqrt{\sum \left[ New(A) - Old(A) \right]^2}$$

The residual is then compared to a threshold to determine if convergence has occurred. In this implementation, a threshold of .001 was used to determine convergence.

## Comparison of Authorities/Hubs and Vector Space Model

In the Vector Space model, similarity is based entirely on the content of a page. If a page contains many of the query terms, it is likely to be scored as very important. This poses a problem in that some very important pages might not contain the query terms but are still highly relevant to the query itself. Authorities/Hubs tries to solve this problem by first finding relevant results based on content, and then by expanding the results and ranking them based on link and citation popularity. Although the Authorities/Hubs model allows relevant pages, which don’t contain the query terms, to appear in the final results, it also causes a problem. When expanding the base set, pages pointed to by the root set are added to the base set. Sites often use a similar look and feel across their pages, many of which include a link to the site homepage on each one of its pages. This results in the homepage having a very high authority score regardless of the query as is the case with the crawled ASU pages. To help correct this problem, characters on either side of the link can be evaluated. If these characters contain any of the query terms, it can be assumed that the link is more important and thus can be assigned a higher weight. This can help to prevent irrelevant pages like homepages from obtaining such a high authority score.

## Effect of Changing the Root Size

When the root size is changed in the Authorities/Hubs calculation, there is a noticeable affect on the relevance of the results. The query “Software Engineering” was tested using root set sizes *k* of 10, 20, and 100. These results can be found in Appendix A. When *k* is 10 the top results appear to be pretty relevant in that the syllabi for 4 courses, each of which has “Software Engineering” in the course title, appear. If *k* is set to 20, the number of pages that do not contain the query terms relative to the pages that do, increase. This means that the chance of base set containing totally irrelevant pages has increased. The results from using a root size of 20 don’t appear to be as relevant as the ones when the root size was 10. Although the syllabi from the previously mentioned courses still appear in the top 10, they appear lower in the rankings. In addition, a new page appears as the third ranked result and appears to be related to a campus tour making it irrelevant. Finally, *k* was set to 100 drastically increasing the number of irrelevant pages in the top ten results. When looking at the results, notice that most of the results from the test with root size 10 no longer appear at all. In addition, because the base set, adjacency matrix, and authorities/hubs scores are all calculated after the user enters the query, the time taken to process the query when *k* is equal to 100 is almost seven times longer than the when *k* was equal to 10. This response time makes a root size of 100 a bad choice for web based search applications. Although it takes more time, the increase in time is not linear; it is actually somewhat logarithmic.

## Authorities/Hubs Relevance

When computing the rank of pages using the Authorities/Hubs algorithm, the authority pages are shown to the user once the scores converge. This is because the authority pages are thought to be the most relevant. In this implementation, the top 10 authority and hub results were computed for several queries. These results can be seen in the attached outputs document. In contrast to the algorithm, the top 10 hub pages actually appear to be more relevant than the top 10 authority pages. Many of the authority results consist of the homepages for ASU, ASU East, ASU West, or the copyright page. This is due to the problem discussed earlier where in many pages have a similar link to a homepage or copyright notification which in turn causes these pages to have high authority scores even though they are irrelevant.

## Conclusion

The Authorities/Hubs ranking algorithm can be useful in that it allows relevant pages which don’t contain the query terms to be included in the top results. This is done by allowing pages to propagate their scores to one another based on their links. This same feature of Authorities/Hubs causes irrelevant pages to appear in the top results regardless of the query. This is caused by links that are added to every page often the result of a similar page template. The pages pointed to by these links wind up getting a high authority score and causing the quality of the results to decline.

# Appendix A

**Authorities/Hubs Ranking Results (k=10)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.421 seconds.

Rank    Hub                   URL
----    ----------            ---
1       0.9881349205674054    www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
2       0.07343406386531984   www.asu.edu%%aad%%catalogs%%general%%engineering.html
3       0.06744458032025324   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
4       0.06744458032025324   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
5       0.06744458032025324   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
6       0.06744458032025324   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
7       2.1930199697497134E-4 www.eas.asu.edu%%CEAS%%academia%%index.shtml
8       2.1930199697497134E-4 www.eas.asu.edu%%CEAS%%news%%about.shtml
9       2.1930199697497134E-4 www.eas.asu.edu%%CEAS%%resources%%asures.shtml
10      2.1930199697497134E-4 www.eas.asu.edu%%CEAS%%students%%careers.shtml
```

**Authorities/Hubs Ranking Results (k=20)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.672 seconds.

Rank    Hub                   URL
----    ----------            ---
1       0.8081207396263159    www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
2       0.5273146856075411    www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
3       0.11987585565539802   www.asu.edu%%tour%%main%%ec.html
4       0.06703378397404489   www.asu.edu%%aad%%catalogs%%general%%ceas.html
5       0.06160393891158453   www.asu.edu%%aad%%catalogs%%general%%engineering.html
6       0.056438033863686836  www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
7       0.056438033863686836  www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
8       0.056438033863686836  www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
9       0.056438033863686836  www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
10      0.056438033863686836  www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl438.html
```

**Authorities/Hubs Ranking Results (k=100)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 2.891 seconds.

Rank    Hub                   URL
----    ----------            ---
1       0.937336182416174     www.eas.asu.edu%%~csedept%%people%%faculty%%faculty.shtml
2       0.1622860417123951    www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
3       0.09805107807046617   www.asu.edu%%asunews%%academics%%online_eng_020404.htm
4       0.09805107807046617   www.asu.edu%%asunews%%university%%minorityscholarship_042303.htm
5       0.06742846267356956   www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
6       0.06726664872322986   www.eas.asu.edu%%CEAS%%resources%%ceasres.shtml
7       0.0671788096399211    www.eas.asu.edu%%CEAS%%depts%%index.shtml
8       0.06712890405932209   www.eas.asu.edu%%CEAS%%students%%careers.shtml
9       0.06710381456596606   www.eas.asu.edu%%CEAS%%resources%%index.shtml
10      0.05896955975120813   www.fulton.asu.edu%%~eee%%Grad%%me.html
```

# CSE 494 Project B – Task 4

# Garrett Wolf

## Introduction

The main purpose of this project task was for us to implement the PageRank + Vector Space algorithm. Once implemented, we were asked to vary the dampening factor *c* and the weight *w* then observe the results. We were provided with a link matrix from the crawled files and code allowing us to get the inbound and outbound links for each of these files. The link matrix was used to implement the PageRank + Vector Space algorithm and integrate it into the application from Project A. This write-up will provide a background on the PageRank + Vector Space ranking algorithm, a comparison of the Authorities/Hubs results to those of the PageRank + Vector Space algorithm, an analysis of the affects on the results from changing the values of *c* and *w*, and a discussion on whether or not PageRank computation converges.

## Page Rank + Vector Space Algorithm Background

The PageRank model is based on the idea that a page with a lot of incoming links must be an important page. In this model, the importance of a page is the probability that a web surfer will end up on that page. To compute the PageRank score of a page, a fraction of the PageRank scores from each of the pages pointing to the page is summed. In this implementation, a page's PageRank score and Vector Space Similarity score are combined to get the final score for the document. To accomplish this, the PageRank score of each page must be precomputed. The provided link matrix file is used to construct a stochastic transition matrix. First, each of the *N* entries in the transition matrix *M* is set to zero. Next, for each page *p*, all the outbound links are obtained from the link matrix file. If there are no outbound links from *p*, *M* is taken and every entry in the column corresponding to *p* is set equal to 1/*N*. In this case, *p* is called a rank sink because it cannot propagate its PageRank to any other pages and thus causes the loss of total ranks. Assigning each entry in the column a value of 1/*N* is based on the random-walk model. The random-walk model says that at any time, there is 1/*N* probability that the surfer can type in a new address and jump to any one of the pages. If page *p* does have outbound links, each entry in column *p* corresponding to one of the linked pages is set equal to 1/*p-out* where *p-out* is the total number of outbound links from *p*. The basic idea is that the PageRank of *p* is equally divided and propagated to each of the pages linked to by *p*. Next, each entry in *M* is multiplied by *c*, the dampening factor used conceptually add a link from each page to every other page including itself. This is another example of the random-walk model. After multiplying each entry by *c*, (1 – *c*)/*N* is added to each entry which assigns the small probabilities that a user randomly jumps to a new page. Now the stochastic transition matrix is complete and each column in the new matrix *M** should sum to one. When creating this matrix, the following formula was used

$$M^* = c(M + Z) + (1 - c)K$$

Here *c* is the dampening factor which is between 0 and 1 but is usually set rather close to 1. *M* is the transition matrix PageRanks and *Z* is the matrix of PageRanks for the rank sink pages. In this implementation, *M* and *Z* were combined from the beginning to make better use of memory resources. *K* is a matrix where every entry is equal to 1/*N*. Again in this implementation, *K* is not explicitly defined. Instead, (1 – *c)*/*N* was added to each entry in the *M*+*Z* matrix to save resources. By using *M** instead of *M*, pages without links to one another and rank sink pages are accounted for and *M** satisfies the conditions that are required to guarantee convergence. Using *M**, the PageRank scores can now be computed. First, each page is assigned a starting PageRank equal to 1/*N*. Next, the PageRanks are computed iteratively until a residual is less than the defined threshold of .001. When this condition is met, the PageRanks have converged. For each page *p,* the new PageRank of *p* is incremented by the PageRank of *q* times element *q,p* in *M** for each page *q*. Once complete, the residual is calculated as

$$residual = \sqrt{\sum \left[ New(PR) - Old(PR) \right]^2}$$

where *New(PR)* and *Old(PR)* equal the new PageRank and previous PageRanks respectively. If the residual is below the threshold, the PageRanks are normalized; otherwise, the PageRanks are computed again until it is. Now that the PageRanks are computed for each page, the user's query is evaluated using the Vector Space Similarity algorithm. Finally, the PageRank + VectorSpace scores are computed using the formula

$$score = w * PageRank + (1 - w) * VectorSpace$$

where *w* is between 0 and 1. The results are then ranked according to this score and displayed to the user.

## Comparison of Authorities/Hubs and PageRank + Vector Space

The idea behind the Authorities/Hubs and PageRank + Vector Space ranking algorithms is similar. Both algorithms differ from pure Vector Space Similarity in that they use links between pages to propagate rank. Both require a matrix depicting the links between the pages in their calculations. Although both algorithms iteratively compute the ranks until a residual falls below a threshold and convergence occurs, the performance requirements are quite different. PageRanks are computed offline and therefore have no strict time requirements whereas Authorities/Hubs are calculated after the user enters a query and thus have a much smaller window for computation. Because of this, Authorities/Hubs is not suited for ranking a large corpus. The Authorities/Hubs algorithm suffers from the impact of bridges in which two disconnected sub graphs of pages are connected by the addition of a page with links to both. This causes the results to be very unstable and makes spamming of the results fairly easy. PageRank on the other hand is harder to spam and provides more relevant results. This is because PageRank calculates part of its final score using the Vector Space Similarity score whereas Authorities/Hubs only calculates it based on popularity. In Appendix A, the results from the query "Fall semester" are shown for both Authorities/Hubs and PageRank + Vector Space. In the Authorities/Hubs model, the top authority results are usually the ones returned to the user. In this case, the top authority result is relevant but all the rest are not. Although the top authority has a high score relative to the other authorities, the results show very poor recall in that only one relevant document was returned. The hub results are slightly better but assign the same score to each of the top 10 results making it hard to choose between them.  The PageRank + Vector Space results appear to be relevant in that they include registrar, prereg, admissions, and catalog pages.  In addition, the PageRank + Vector Space results were returned in approximately one fourth the time that it took for the Authorities/Hubs results.

## Effect of Changing the Dampening Factor

By changing the dampening factor *c*, more or less probability can be given to a user randomly jumping to a page.  For different types of users, this might be useful in that experienced users may be familiar with many pages and are thus more likely to jump to one of them than is a person who is using the web for the first time.  In Appendix B the results of the query "Software Engineering" can be seen for three different values of *c* namely, .7, .8, and .9.  It is obvious that the top results are virtually identical just slightly reordered.  If *c* is set to .2 we are basically saying that there is an 80% chance that the user with randomly jump to a new page.  This is probably not a very realistic probability but it still has very little affect on the top 10 results.  One difference though is that the scores for the results when *c* is .2 are almost double the other trials.  This is because the most pages don't link to most other pages so most of the values in *M** are much higher than in the other cases.

## Effect of Changing the PageRank Weight

By changing the PageRank weight *w*, more or less weight can be given to PageRank or Vector Space Similarity.  In Appendix C the results of the query "Software Engineering" can be seen for *w* values of .2, .5, and .8.  Using *w* equals .5 as the baseline, we can see that when *w* is increased, many of the syllabi disappear from the top 10.  This is because those classes contain the query term in their title and thus are have high Vector Space Similarity scores.  By increasing *w,* a smaller weight Vector Space weight is given to them and thus they do not appear.  In contrast, if we lower the value of *w*, we see them reappear and very close to the top.  Therefore increasing *w* returns results that are more important based on link structure whereas decreasing *w* returns results that are more important based on content.

## PageRank Convergence

In this implementation, the PageRanks do converge because both requirements for convergence are satisfied.  First, the graph *M** of the pages is aperiodic which is practically guaranteed for the Web.  Second, *M** is irreducible because rank sinks and pages without links to every other page were accounted for by the random-walk model.  For the ranks to converge, there must be a rank vector *R* such that *R* is the eigenvector of matrix *M** with eigenvalue being 1.  Because *M** is stochastic, we know that the principle eigen value is 1 and therefore the ranks do converge.

## Conclusion

The PageRank + Vector Space ranking algorithm appears to provide quality results.  This based on the fact that both link and content analysis are used to rank the results.  PageRank values can be computed offline for each page in the corpus allowing a quick turnaround from the time when the user enters their query.  The PageRank model eliminates some of the problem found in the Authorities/Hubs model by using a random-walk model. The random-walk model is based on the idea that a user can jump to a new page at any time without following the links on their current page. This helps to improve the quality of the results returned to the user.

# Appendix A

**Authorities/Hubs Ranking Results (k=10)**

```
Query: Fall semester
Searching for: fall semester
Search completed in 0.469 seconds.

Rank    Hub                   URL
----    ----------            ---
1       0.14142107180229568   www.west.asu.edu%%chs%%SW%%faculty.htm
2       0.14142107180229568   www.asu.edu%%workingatasu%%academics%%newfaculty.html
3       0.14142107180229568   www.east.asu.edu%%ecollege%%html%%degrees.htm
4       0.14142107180229568   www.east.asu.edu%%ecollege%%appliedbiologicalsciences%%contact.html
5       0.14142107180229568   www.west.asu.edu%%chs%%
6       0.14142107180229568   clasdean.la.asu.edu%%student%%resources%%
7       0.14142107180229568   www.asu.edu%%duas%%cas%%changemajor.htm
8       0.14142107180229568   www.east.asu.edu%%academics%%resources%%
9       0.14142107180229568   www.west.asu.edu%%chs%%Student-services%%
10      0.14142107180229568   www.west.asu.edu%%chs%%grn%%

Rank    Authority             URL
----    ----------            ---
1       0.999997175359144     www.asu.edu%%registrar%%general%%semcal.html
2       0.0014459361664669042 www.asu.edu
3       5.734773081600025E-4  www.west.asu.edu%%
4       4.968715296878437E-4  www.east.asu.edu%%
5       4.8789094704226584E-4 www.asu.edu%%
6       4.112851685701071E-4  www.asu.edu%%xed%%
7       3.06343436208861E-4   www.asu.edu%%copyright%%
8       2.6089999487787023E-4 www.asu.edu%%copyright
9       2.519194122322923E-4  www.west.asu.edu
10      2.519194122322923E-4  www.west.asu.edu%%ams%%profiles
```

**Page Rank + Vector Space Ranking Results (w=.5 and c=.8)**

```
Query: Fall semester
Searching for: fall semester
Search completed in 0.125 seconds.

Rank    Page Rank             URL
----    ----------            ---
1       0.39035930269204766   www.asu.edu%%registrar%%general%%semcal.html
2       0.37996142215349643   coe.asu.edu%%oss%%prereg%%index.php
3       0.3381568454853596    www.eas.asu.edu%%~cse355%%cse355B%%index.html
4       0.3306846164412837    geology.asu.edu%%courses%%index.html
5       0.29683418433529163   www.west.asu.edu%%ams%%Degrees%%2003_2004%%Hisminor.htm
6       0.29400629005094303   www.west.asu.edu%%amerstud%%Degrees%%2003-2004%%Hisminor.htm
7       0.29262649970425      www.asu.edu%%admissions%%steps%%application.html
8       0.286479359750223     www.asu.edu%%aad%%catalogs%%catalog-calendars.html
9       0.2804236386468725    www.asu.edu%%aad%%catalogs%%general%%asue-business-admin.html
10      0.277195840395625     www.asu.edu%%lib%%music%%hours%%index.html
```

# Appendix B

**Page Rank + Vector Space Ranking Results (w=.5 and c=.7)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.172 seconds.

Rank    Page Rank              URL
----    ----------             ---
1       0.5383498671281276     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
2       0.49039521269520026    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
3       0.4897620883284299     www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
4       0.4864815679474055     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
5       0.47087448111551367    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
6       0.44180395478251233    www.east.asu.edu%%explore%%spotlight1.html
7       0.43113747899583793    www.asu.edu%%aad%%catalogs%%general%%engineering.html
8       0.41889900675684444    www.asu.edu%%aad%%catalogs%%general%%ceas.html
9       0.4184297517101169     www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
10      0.4093625303653263     www.fulton.asu.edu%%%7Ecivil%%UG_What.htm
```

**Page Rank + Vector Space Ranking Results (w=.5 and c=.8)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.125 seconds.

Rank    Page Rank              URL
----    ----------             ---
1       0.454417795203918      www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
2       0.40646314077099066    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
3       0.4025494960231959     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
4       0.3990153130589214     www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
5       0.38694240919130407    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
6       0.35076624924185534    www.east.asu.edu%%explore%%spotlight1.html
7       0.34142652814439745    www.asu.edu%%aad%%catalogs%%general%%engineering.html
8       0.32842815628916244    www.asu.edu%%aad%%catalogs%%general%%ceas.html
9       0.3275674626390339     www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
10      0.32405377578616623    www.eas.asu.edu%%~cse461%%top.html
```

**Page Rank + Vector Space Ranking Results (w=.5 and c=.9)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.125 seconds.

Rank    Page Rank              URL
----    ----------             ---
1       0.4097297715771136     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
2       0.36177511714418625    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
3       0.35786147239639154    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
4       0.3502044287024231     www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
5       0.3422543855644997     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
6       0.30156599754336133    www.east.asu.edu%%explore%%spotlight1.html
7       0.29899112720668797    www.eas.asu.edu%%~cse461%%top.html
8       0.29389560883096677    www.asu.edu%%aad%%catalogs%%general%%engineering.html
9       0.2791307788220074     www.asu.edu%%aad%%catalogs%%general%%ceas.html
10      0.2786639228622314     www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
```

**Page Rank + Vector Space Ranking Results (w=.5 and c=.2)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.157 seconds.

Rank    Page Rank             URL
----    ----------            ---
1       0.806922674440139     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
2       0.7589680200072116    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
3       0.755054375259417     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
4       0.7394472884275252    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
5       0.7012549844084469    www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
6       0.6832762698519277    www.eas.asu.edu%%~cse461%%top.html
7       0.656479603696384     www.asu.edu%%aad%%catalogs%%general%%engineering.html
8       0.6533798801327013    www.east.asu.edu%%explore%%spotlight1.html
9       0.6450731732071778    acims.eas.asu.edu%%EVENTS%%DLS03%%synopsis.htm
10      0.644556320818963     www.asu.edu%%aad%%catalogs%%general%%ceas.html
```

# Appendix C

**Page Rank + Vector Space Ranking Results (w=.5 and c=.8)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.188 seconds.

Rank    Page Rank             URL
----    ----------            ---
1       0.454417795203918     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
2       0.40646314077099066   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
3       0.4025494960231959    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
4       0.3990153130589214    www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
5       0.38694240919130407   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
6       0.35076624924185534   www.east.asu.edu%%explore%%spotlight1.html
7       0.34142652814439745   www.asu.edu%%aad%%catalogs%%general%%engineering.html
8       0.32842815628916244   www.asu.edu%%aad%%catalogs%%general%%ceas.html
9       0.3275674626390339    www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
10      0.32405377578616623   www.eas.asu.edu%%~cse461%%top.html
```

**Page Rank + Vector Space Ranking Results (w=.8 and c=.8)**

```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.125 seconds.

Rank    Page Rank             URL
----    ----------            ---
1       0.39678147660486995   www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
2       0.37703309962845893   www.east.asu.edu%%explore%%spotlight1.html
3       0.3695434409443617    www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
4       0.3605216483047956    www.eas.asu.edu%%~csedept%%people%%faculty%%faculty.shtml
5       0.3596281603739384    www.asu.edu%%feature%%computer%%protect.html
6       0.358624219998834     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
7       0.35836987759610717   www.asu.edu%%aad%%catalogs%%general%%engineering.html
8       0.3576404323705189    www.eas.asu.edu%%CEAS%%resources%%ceasres.shtml
9       0.3563519070125001    www.fulton.asu.edu%%%7Ecivil%%Undergrad.htm
10      0.35633787060085886   www.eas.asu.edu%%CEAS%%resources%%index.shtml
```

**Page Rank + Vector Space Ranking Results (w=.2 and c=.8)**

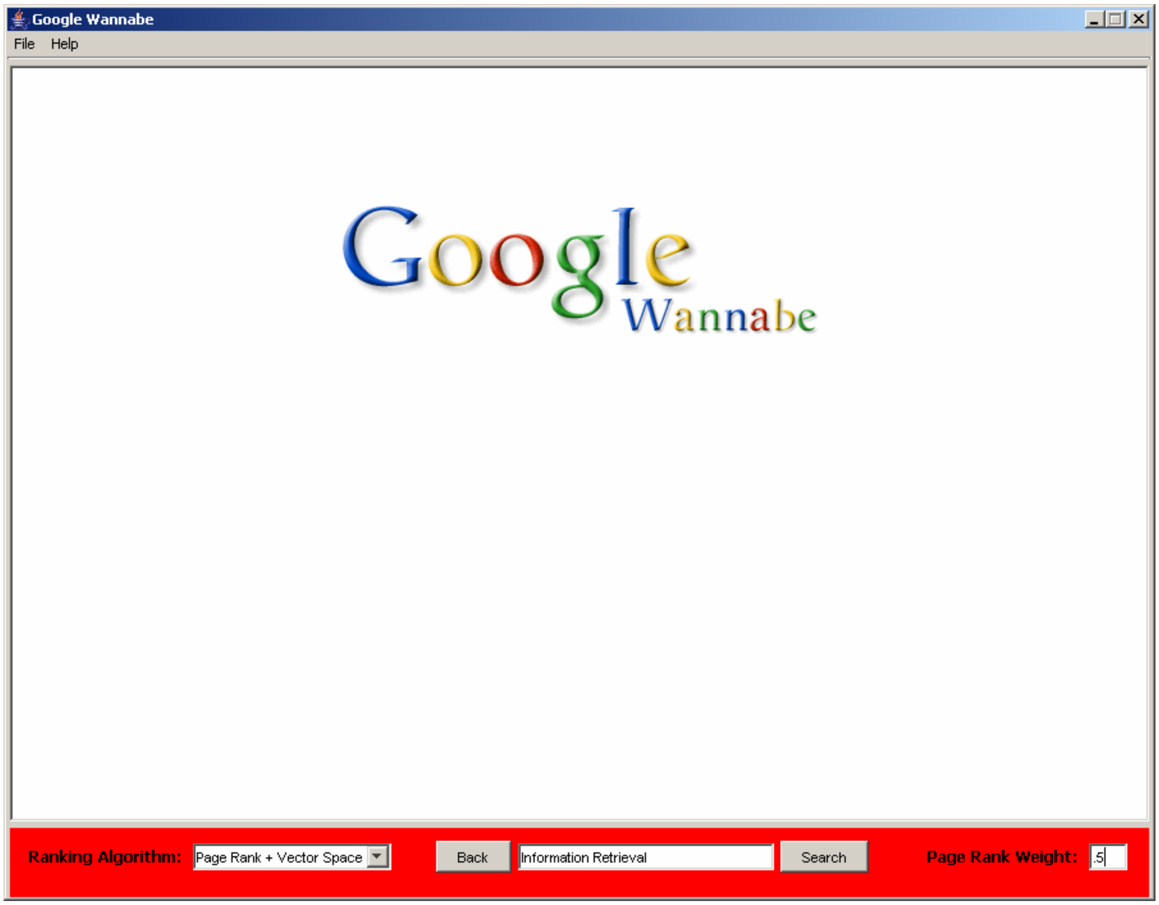
```
Query: Software Engineering
Searching for: software engineering
Search completed in 0.11 seconds.

Rank    Page Rank             URL
----    ----------            ---
1       0.550211370409002     www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
2       0.4734839233163182    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
3       0.4696279142599183    www.eas.asu.edu%%~cse461%%top.html
4       0.4672220917198467    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
5       0.44225075278881976   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
6       0.40124914951297286   www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
7       0.3832853627448322    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl562.html
8       0.34750996847303406   www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl564.html
9       0.3342055479738644    www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl563.html
10      0.3244993988552518    www.east.asu.edu%%explore%%spotlight1.html
```

# CSE 494 Project B – Extra Credit

# Garrett Wolf

## Introduction

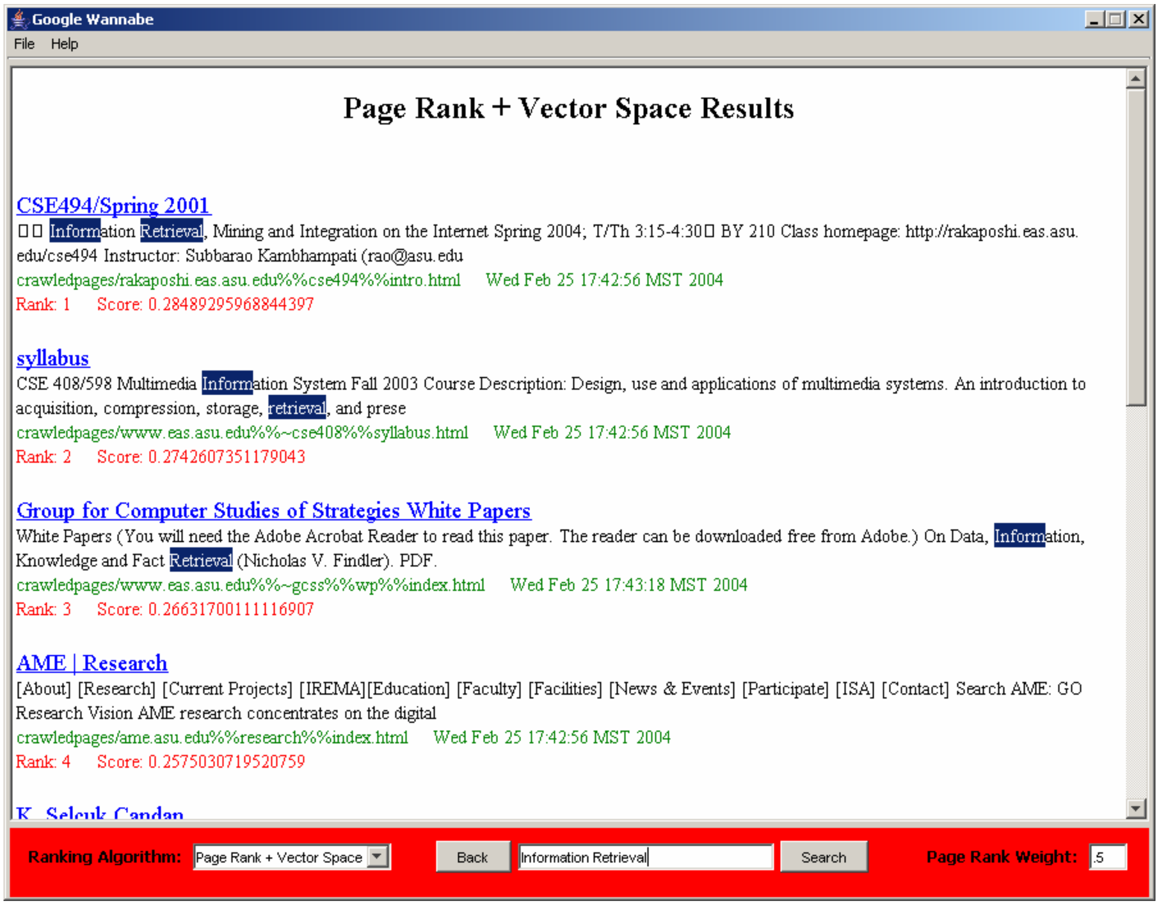
The main purpose of this project task was to create GUI interface to access the underlying search engine. This implementation was built using the javax.swing package. Because I did not have access to a web server, I created a simple java web browser to view the query results. The user can select a ranking algorithm, set appropriate values such as *k*, *w*, and *c*, and enter a search query. When the search button is clicked, the html content results are generated and displayed to the user. Each result includes the page title, summary, url, modified date, rank, and score. In addition, each query term is highlighted on the results page. The user can then click on one of the links to open the document in the browser window. From there, the user can click other links on that page and be taken out to the web where any of those pages can be displayed as well.



Ranking Algorithm: Page Rank + Vector Space

Information Retrieval

Page Rank Weight: .5



## Page Rank + Vector Space Results

CSE494/Spring 2001

□□ Information Retrieval, Mining and Integration on the Internet Spring 2004; T/Th 3:15-4:30□ BY 210 Class homepage: http://rakaposhi.eas.asu.edu/cse494 Instructor: Subbarao Kambhampati (rao@asu.edu

crawledpages/rakaposhi.eas.asu.edu%%cse494%%intro.html     Wed Feb 25 17:42:56 MST 2004

Rank: 1     Score: 0.28489295968844397

syllabus

CSE 408/598 Multimedia Information System Fall 2003 Course Description: Design, use and applications of multimedia systems. An introduction to acquisition, compression, storage, retrieval, and prese

crawledpages/www.eas.asu.edu%%~cse408%%syllabus.html     Wed Feb 25 17:42:56 MST 2004

Rank: 2     Score: 0.2742607351179043

Group for Computer Studies of Strategies White Papers

White Papers (You will need the Adobe Acrobat Reader to read this paper. The reader can be downloaded free from Adobe.) On Data, Information, Knowledge and Fact Retrieval (Nicholas V. Findler). PDF.

crawledpages/www.eas.asu.edu%%~gcss%%wp%%index.html     Wed Feb 25 17:43:18 MST 2004

Rank: 3     Score: 0.26631700111116907

AME | Research

[About] [Research] [Current Projects] [IREMA][Education] [Faculty] [Facilities] [News & Events] [Participate] [ISA] [Contact] Search AME: GO Research Vision AME research concentrates on the digital

crawledpages/ame.asu.edu%%research%%index.html     Wed Feb 25 17:42:56 MST 2004

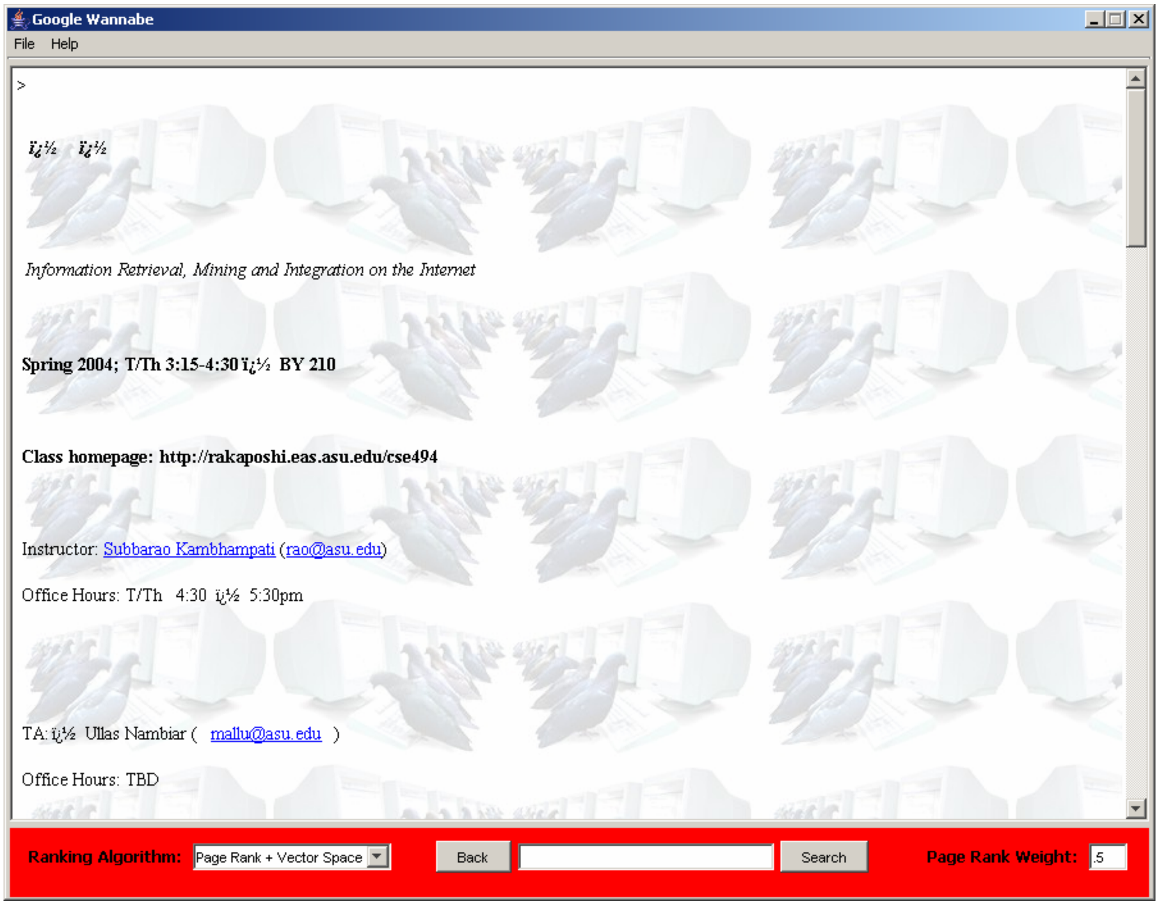
Rank: 4     Score: 0.2575030719520759

K. Selcuk Candan

Ranking Algorithm: Page Rank + Vector Space

Information Retrieval

Page Rank Weight: .5



>

**ï¿½ ï¿½**

*Information Retrieval, Mining and Integration on the Internet*

**Spring 2004; T/Th 3:15-4:30 ï¿½ BY 210**

**Class homepage: http://rakaposhi.eas.asu.edu/cse494**

Instructor: Subbarao Kambhampati (rao@asu.edu)

Office Hours: T/Th 4:30 ï¿½ 5:30pm

TA: ï¿½ Ullas Nambiar ( mallu@asu.edu )

Office Hours: TBD

Ranking Algorithm: Page Rank + Vector Space

Page Rank Weight: .5