Project B Report (Authority Hub)

This part of report contains details about Authority Hub computation. First the Authority/Hub computation algorithm is discussed. Then comparison of Authority/Hub results with Vector Space similarity results is presented. After that the effect of varying Root Set size is discussed.

Authority Hub Algorithm

A page referenced by many important pages is more important (authority) and a page that references a lot of important pages is also more important (Hub). The algorithm starts by running Vector Space algorithm. Top k documents are taken from this result. They form the root set. Root set is expanded into base set by taking documents that are linked by or linking to at least one document of root set. Only limited no. of documents are taken that point to the documents in the root set as no. of documents can be large that point to a document in root set.

This algorithm is started by constructing adjacency matrix A for pages in base set. Each entry Aij in matrix A contains 1 if there is a link from document i to document j in base set. Otherwise it is set to 0. Authority and hub values of each document are set to one. Computation is iterative. In each iteration authority value of a document is computed by adding hub values of all the documents that point to the given document. After computing authority value of each document, hub value of each document is computed. Hub value of a document is the sum of authority value of each document the current document points to. After that authority and hub values of each document are normalized by dividing each authority value by the length of authority vector and same is done for hub value of each document.

At the end of any iteration, convergence is checked. Threshold is set to 0.00001. Difference between authority values of any iteration and its previous one is compared to this threshold to find convergence. The algorithm is given below.

```
AuthorityHub(Query Q){
        Find relevant documents by using Vector Space similarity
        Construct root set by using top k documents
        Construct Hash Tables pointsTo and pointedBy for storing forward and backward
        links respectively

        For each page p in root set {
                If p has forward links {
                        pointsTo.put (p, list of all forward links from p)
                }

                If p has back links {
```

```
                pointedBy.put (p, list of all backward links upto 50)
            }
        }
        For each page p in baseSet and not in root set {
            If p has forward links {
                pointsTo.put (p, list of all forward links from p)
            }
        }

        authorityValues[1..N] = 0
        hubValues[1..N] = 1

        For 50 iterations or until convergence {
            For each page p in pointsTo.keys{
                For each page q in pointsTo[p] {
                    authorityValues[q] += hubValues[p]
                }
            }
            For each page p in pontedBy.keys {
                For each page q in pointedBy[p] {
                    hubValues[q] += authorityValues[p]
                }
            }
            Normalize authorityValues and hubValues
        }
        Sort authorityValues and hubValues
}
```

Here algorithm stops at 50 iterations or when the authority and hub values converge whichever occurs first. If there are N documents in total the Vector Space results are computed in O (N log N).

If there are k documents in root set and there are F forward links and B backward links for each page then time complexity of expanding root set to base set is
$O (k * (F + B) + k * (F + B) * F) = O (F^2)$

To compute authority and hub values the time complexity is

$O (50 * ((k + k * (F + B)) * F + (k + k * (F + B)) * B))$
$= O (50 * (kF + kF^2 + 2kBF + kB^2 + kB))$
$= O (F^2 + B^2)$

To sort these pages time required is
$O ((k + kF + kB) \log (k + kF + kB)) = O ((F + B) \log (F + B))$

Total time complexity is

$O (N \log N + F^2 + F^2 + B^2 + (F + B) \log (F + B))$
$= O (N \log N + F^2 + B^2)$

Space required to store forward and backward links is of order
$O (k * B + k (F + B) * F) = O (F^2)$
And to store authority and hub values space required is of $O (k * (F + B)) = O (F + B)$
Total space complexity is of $O (F^2)$.

Comparison of Vector Space and Authority/Hub results

Vector space returns documents that are highly similar in content with the query. It is possible that though the document is not similar with query in terms of content but it is relevant to the query. Authority/Hub computation takes importance of the document into account.

But Authority/Hub computation has a major drawback. If base set of given query contains a page that is highly important but not relevant to the query then also that page ends up having high authority value. Like in this experiment www.asu.edu ends up being top authority for many queries. To overcome this we can take Vector Space similarity of anchor text and query and add only those documents that are similar to query.

Effect of varying the size of root set

In the experiment root set values of 10 and 20 is used to measure the effect of changing size. As the root set size changes the base set size also changes almost proportionally. And as the root set size changes the results also change. Following are the results for "transcripts" query with root set size 10 and 20. Convergence threshold used to stop A/H computation is 0.00001.

Pages before extension of set: 10
Pages after extension of set: 122

Authority Values	URL
0.387307	www.asu.edu
0.28527978	www.asu.edu%%copyright%%
0.2772719	www.asu.edu%%privacy%%
0.24796544	www.asu.edu%%registrar%%transcripts%%index.html
0.24680589	www.asu.edu%%interactive
0.23680107	www.asu.edu%%registrar%%graduation%%index.html

0.23549329 www.asu.edu%%registrar%%registration%%index.html
0.23549329 www.asu.edu%%vpsa%%
0.23549329 www.asu.edu%%registrar%%forms%%index.html
0.23549329 www.asu.edu%%registrar%%general%%contacts.html

Pages before extension of set: 20
Pages after extension of set: 235

Authority Values	URL
0.3556297	www.asu.edu
0.23042277	www.asu.edu%%copyright%%
0.21872695	www.asu.edu%%privacy%%
0.18825449	www.west.asu.edu%%
0.18720019	www.asu.edu%%
0.16406149	www.east.asu.edu%%
0.15752876	www.asu.edu%%xed%%
0.14396375	www.east.asu.edu%%admissions%%
0.13903569	www.east.asu.edu%%students%%
0.13903569	www.east.asu.edu%%about%%welcome%%

For this query the results are changed dramatically except for the top authority because it is the most important page in the given corpus. And many pages contain link to it so its authority value is going to be stable. For rest of pages, on increasing root set size more relevant documents are included and those pages may contain links to pages that are more important link wise. So results change dramatically.

For some queries like “parking decal” varying the size of root set doesn’t change document ranking. Because on increasing root set size newly added documents are less relevant and may be less important also link wise. So results are not changed much.

Pages before extension of set: 10
Pages after extension of set: 105

Authority Values	URL
0.16018932	www.asu.edu%%
0.1579589	www.asu.edu%%copyright%%
0.15553312	www.asu.edu%%privacy%%
0.15511373	www.east.asu.edu%%
0.15432149	www.west.asu.edu%%
0.15404968	www.asu.edu%%xed%%
0.15285926	www.east.asu.edu%%admissions%%
0.15276563	www.east.asu.edu%%about%%weather%%

0.15256576 www.east.asu.edu%%teaching%%policy%%
0.15256576 www.east.asu.edu%%about%%units%%

Pages before extension of set: 20
Pages after extension of set: 142

Authority Values	URL
0.16399951	www.asu.edu%%
0.15873112	www.asu.edu%%copyright%%
0.15812725	www.east.asu.edu%%
0.15732318	www.west.asu.edu%%
0.15548882	www.asu.edu%%privacy%%
0.15406623	www.asu.edu%%xed%%
0.15247121	www.east.asu.edu%%admissions%%
0.15237822	www.east.asu.edu%%about%%weather%%
0.15217654	www.east.asu.edu%%about%%personnel%%
0.15217654	www.east.asu.edu%%working%%policy%%

Relevance of Authority/Hub pages

The result of this experiment shows that Hub values are more relevant than authority values. Following are the top 10 authority and hub results for query “transcripts”.

Pages before extension of set: 15
Pages after extension of set: 192

Hub Values	URL
0.2554889	www.east.asu.edu%%about%%personnel%%
0.17669402	www.east.asu.edu%%contact%%
0.15907481	www.east.asu.edu%%admissions%%requirements%%transfer.html
0.15881015	www.east.asu.edu%%admissions%%requirements%%international.html
0.15880999	www.east.asu.edu%%admissions%%requirements%%nondegree.html
0.15880999	www.east.asu.edu%%admissions%%requirements%%other.html
0.15338701	www.east.asu.edu%%admissions%%requirements%%
0.15327148	www.east.asu.edu%%index%%
0.15294233	www.east.asu.edu%%admissions%%requirements%%graduate.html
0.1525577	www.east.asu.edu%%working%%resources%%

Authority Values	URL
0.24984981	www.asu.edu

0.19983836 www.asu.edu%%copyright%%
0.19498207 www.asu.edu%%privacy%%
0.17806925 www.asu.edu%%
0.17078678 www.west.asu.edu%%
0.16477461 www.east.asu.edu%%
0.15672958 www.asu.edu%%xed%%
0.15177532 www.east.asu.edu%%admissions%%
0.15099958 www.east.asu.edu%%working%%
0.15099958 www.east.asu.edu%%about%%units%%

Here www.asu.edu comes as top authority instead of having very less relevance to given query. Because it is important page in entire corpus and it is pointed by many pages and comes in base set.

Hub values represent relevant results because general high authority pages (e.g. www.asu.edu) are pointed by many pages. Authority value of this kind of authority page is spread equally to all pages. In this case, top hub pages are those that contain links to authorities which are specific to given query. So it gives results more relevant to query.

Project B Report (PageRank)

This part of report contains details about PageRank implementation. First the page rank algorithm is discussed. Then comparison of Authority/Hub results with PageRank is presented. After that effect of varying PageRank computation weight and effect of varying damping factor are discussed.

PageRank Algorithm

PageRank algorithm takes into account the popularity of a page. i.e., what is the probability that the surfer will be on this particular page. If a page has links to n pages then the probability that surfer follows any one of this links is 1/n. These probabilities are stored in adjacency matrix. Where each entry Aij represents probability of following page i from page j. So each column must add up to 1. For pages having back links but no forward links, probability of jumping from that page to any page is assumed to be equal (1/n). This assumption solves problem of sink node. It is also possible that surfer follows any page at random from a given page with a low probability.

So the Matrix Looks Like,

$M = C * (A + Z) + (1-C) * K$

Where, A is the stochastic matrix containing probabilities of following from any page to any page. Each entry Zij of Z contains 0 if there is at least one out link from j. Remaining entries contain 1/n (where n = total no. of pages). K represents surfer's random behavior. Each entry of K has value 1/n (where n = total no. of pages). C is a damping factor. It represents the probability with which random surfer will follow links on current page. Its value ranges from 0 to 1 both inclusive.

PageRank of different pages is the principle eigen vector of matrix M. We can compute PageRank by multiplying matrix M with initial PageRank vector iteratively. Initial PageRank of each page is set to 1/n. After PageRank is computed for all the pages, we calculate Vector Space similarity for given query. PageRank and Vector Space similarity values are combined to give final ranking of pages. The algorithm is as follows.

In this algorithm total no. of documents are N.

```
PageRank (C) {
        Construct hash table Links for storing page and its forward links

        For each page p {
                If p has forward links {
                        Links.put (p, list of all forward links)
                        outLinks[p] = no. of forward link from p
                }
        }
```

```
        PageRank[1..N] = 1/N
        Loop until convergence {
                newPageRank[1..N] = 0
                sumSinkPageRank = 0
                totalPageRank = 0

                For each page p {
                        For each page q in Links[p] {
                                newPageRank[q] += damping factor C * PageRank
                        }
                        If p is a sink node {
                                sumSinkPageRank += PageRank[p]
                        }
                        totalPageRank += PageRank[p]
                }

                For each page p {
                newPageRank[p] += ((C*sumSinkPageRank) + ((1 - C) * totalPageRank))
                newPageRank[p] /= totalPageRank
                }

                Normalize newPageRank
                PageRank[1..N] = newPageRank[1..N]
        }
}
```

Here PageRank computation stops at predefined no. of iterations or when the PageRank values converge. Now if average no. of forward links from each page is L then complexity of above algorithm can be calculated as given below. Here maximum no. of iterations is set to 20.

Time complexity to find forward links of each page is O (N * L). So total time complexity is

$$O (N * L + 20 * (N * L + N + N))$$
$$= O (N * L)$$

To combine PageRank with Vector Space Similarity algorithm is as follows.

```
Ranking (w)
{
        Find VectorSpace results for given query
        For each page p in result {
                Score[p] = w * PageRank[p] + (1 – w) * VectorSpaceScore[p]
        }
        Sort scores
}
```

Result from VectorSpace algorithm contains R results. Here time complexity is.

$$O (N \log N) + O (R) + O (R \log R)$$
$$= O (N \log N)$$

Total time complexity becomes O (N * L + N log N).

To store links of each page only space of order N * L is required. And to store PageRank and NewPageRank space of O (N) is required. So total space required is of O(N * L).

Observations

PageRank values are very small except for few pages because sum of PageRanks of different pages should be 1. So as the no. of pages increase, values of PageRanks decrease. In this case Vector Space similarity values dominate the final score. So to combine PageRank with Vector Space Similarity it should be normalized to be in 0 to 1 with inclusive 1. And Vector Space similarity values should also be normalized to be in 0 to 1 with inclusive 1.

Comparison of A/H results and PageRank

PageRank and A/H both take link information into consideration. PageRank is computed before query is given. A/H is computed after finding Vector Space results for a given query.

For pages having no out links, PageRank assigns small probability to jump to any page from such pages. This gives stability to PageRank computation. A/H takes into account only authority hub values to rank final score. PageRank combines Vector Space similarity and PageRank. So PageRank gives documents which are **relevant** as well as **important.**

Following are the results for “Software Engineering” query from A/H computation and PageRank.

Top 10 Hub pages

Hub Values	URL
0.2642647	www.eas.asu.edu%%CEAS%%misc%%contact.shtml
0.21671925	www.eas.asu.edu%%CEAS%%alumni%%advisement.shtml
0.19294657	www.eas.asu.edu%%CEAS%%students%%scholarships.shtml
0.16917388	www.eas.asu.edu%%CEAS%%news%%bios.shtml
0.16917388	www.eas.asu.edu%%CEAS%%news%%logos.shtml
0.15770322	www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml
0.15760438	www.eas.asu.edu%%CEAS%%depts%%index.shtml
0.15741013	www.eas.asu.edu%%CEAS%%resources%%ceasres.shtml
0.1573971	www.eas.asu.edu%%CEAS%%resources%%asures.shtml
0.15736662	www.eas.asu.edu%%CEAS%%academia%%index.shtml

Top 10 Authority pages

Authority Values	URL
0.53029215	www.asu.edu
0.13900854	www.asu.edu%%privacy%%
0.13504975	www.asu.edu%%calendar%%academic.html
0.13417271	cpd.asu.edu%%
0.13357243	www.eas.asu.edu%%~mae%%
0.13357243	www.eas.asu.edu%%CEAS%%students%%index.shtml
0.13357243	www.eas.asu.edu%%CEAS%%industry%%index.shtml
0.13357243	www.eas.asu.edu%%~leaders%%
0.13357243	ceaspub.eas.asu.edu%%ie%%
0.13357243	www.eas.asu.edu%%%7Eindustry%%

Following are the results using PageRank computation.

PageRank + Similarity	URL
0.27958384	www.eas.asu.edu%%~cse461%%top.html
0.27548635	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html
0.2481312	www.eas.asu.edu%%ceas%%depts%%degreeprograms.shtml
0.23434588	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html
0.2268605	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html
0.21324092	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl562.html
0.21071379	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html
0.19338946	www.asu.edu%%aad%%catalogs%%general%%engineering.html
0.18675041	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl564.html
0.17861028	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl563.html

Here it can be seen that PageRank returns more relevant pages to the query than A/H computation. In PageRank result pages are of Software Engineering related courses. In this case PageRank performs better than A/H computation.

Effects of varying PageRank weight

PageRank weight can be varied between 0 and 1. For larger values of this weight, algorithm returns **important** pages. For smaller values of this weight, algorithm returns **relevant** pages.

Results for varying values of w are given below. Query issued is “Multimedia Database”. Damping factor C is set to 0.8. Following are the results for w = 0.75.

PageRank + Similarity	URL
0.11075871	www.public.asu.edu%%~candan%%teaching.htm
0.10167459	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl412.html
0.099051945	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl408.html
0.09790564	www.eas.asu.edu%%~cse412%%nsf_grant.html
0.091776095	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl510.html
0.086132325	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl494db.html
0.0826682	www.public.asu.edu%%~candan%%research.htm
0.081050694	www.east.asu.edu%%ctas%%imt%%html%%git.html
0.08051257	www.public.asu.edu%%~candan%%cv.htm
0.080123305	www.eas.asu.edu%%~cse408%%syllabus.html

Following are the results for w = 0.25.

PageRank + Similarity	URL
0.3266345	www.public.asu.edu%%~candan%%teaching.htm
0.30348024	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl412.html
0.29562432	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl408.html
0.29167244	www.eas.asu.edu%%~cse412%%nsf_grant.html
0.27378476	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl510.html
0.25686547	www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl494db.html
0.24105035	www.east.asu.edu%%ctas%%imt%%html%%git.html
0.24062762	www.public.asu.edu%%~candan%%research.htm
0.237672	www.eas.asu.edu%%~cse408%%syllabus.html
0.23632978	www.eas.asu.edu%%~adood%%idm99workshopreport.html

In both cases results are almost same. Because Vector Space similarity Values are large enough to dominate in top 10 even if small weight is given to Vector Space similarity. Here "**www.public.asu.edu%%~candan%%cv.htm**" is not there in results with higher weight to Vector Space similarity. And it is there in results with high weight for PageRank. This page may not be containing Multimedia Database term more often but it is important for Multimedia Database.

Effect of varying damping factor

Damping factor C is used to change weight to surfer's random behavior and its behavior to follow links given on a given page.

$$PageRank = (C * (A + Z) + (1 - C) * K) * PageRank$$

As the value of C is larger PageRank computation follows links on a page more. As the value of C is smaller it follows random behavior more.

Following are the results for varying damping factor for query “transcripts”. Here w is set to 0.5.

Top 10 results for C = 0.8

PageRank +Similarity	URL
0.3274871	www.asu.edu%%admissions%%steps%%finaltranscripts.html
0.2559525	www.asu.edu%%registrar%%transcripts%%index.html
0.1901339	www.asu.edu%%admissions%%steps%%transcripts.html
0.13120742	www.asu.edu%%admissions%%steps%%dining.html
0.13021575	www.west.asu.edu%%gowest%%stepstrans.htm
0.12574005	www.asu.edu%%admissions%%steps%%campusvisit.html
0.12045341	www.west.asu.edu%%gowest%%steps.htm
0.11538304	www.asu.edu%%admissions%%steps%%parking.html
0.11438408	www.asu.edu%%admissions%%steps%%housing.html
0.11018965	www.asu.edu%%graduate%%admissions%%general_info.html

Top 10 results for C = 0.6

PageRank +Similarity	URL
0.32796827	www.asu.edu%%admissions%%steps%%finaltranscripts.html
0.25427625	www.asu.edu%%registrar%%transcripts%%index.html
0.1906158	www.asu.edu%%admissions%%steps%%transcripts.html
0.13168857	www.asu.edu%%admissions%%steps%%dining.html
0.13056806	www.west.asu.edu%%gowest%%stepstrans.htm
0.12622194	www.asu.edu%%admissions%%steps%%campusvisit.html
0.12080572	www.west.asu.edu%%gowest%%steps.htm
0.11586493	www.asu.edu%%admissions%%steps%%parking.html
0.114865616	www.asu.edu%%admissions%%steps%%housing.html
0.110464476	www.asu.edu%%graduate%%admissions%%general_info.html

Top 10 results for C = 0.4

PageRank +Similarity	URL
0.32874715	www.asu.edu%%admissions%%steps%%finaltranscripts.html
0.25406623	www.asu.edu%%registrar%%transcripts%%index.html
0.19139507	www.asu.edu%%admissions%%steps%%transcripts.html
0.13246748	www.asu.edu%%admissions%%steps%%dining.html
0.13128504	www.west.asu.edu%%gowest%%stepstrans.htm
0.12700121	www.asu.edu%%admissions%%steps%%campusvisit.html

0.121522695	www.west.asu.edu%%gowest%%steps.htm
0.1166442	www.asu.edu%%admissions%%steps%%parking.html
0.115644716	www.asu.edu%%admissions%%steps%%housing.html
0.11114907	www.asu.edu%%graduate%%admissions%%general_info.html

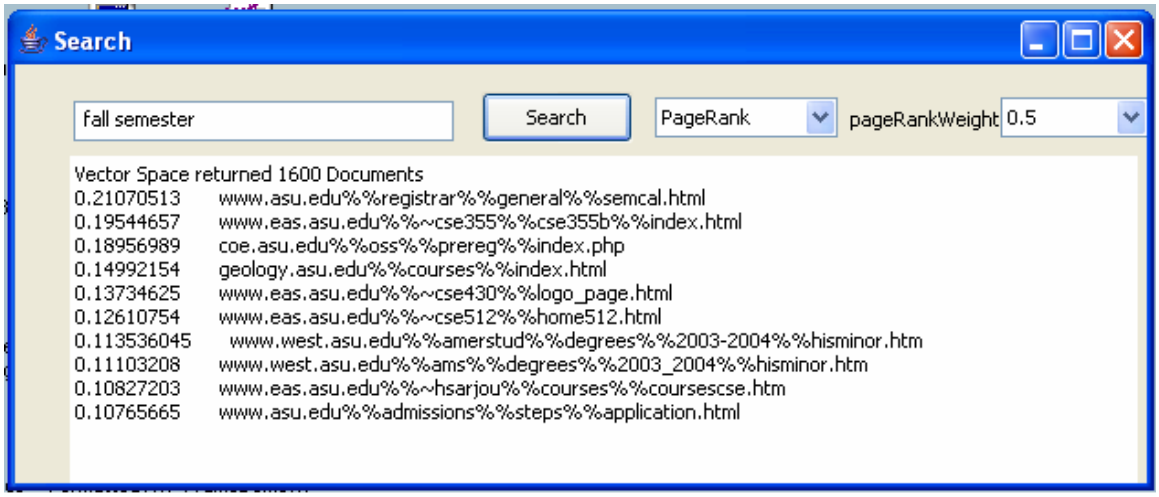
Here ranking of pages doesn't change much on varying damping factor. But the computation becomes faster on decreasing value of C.

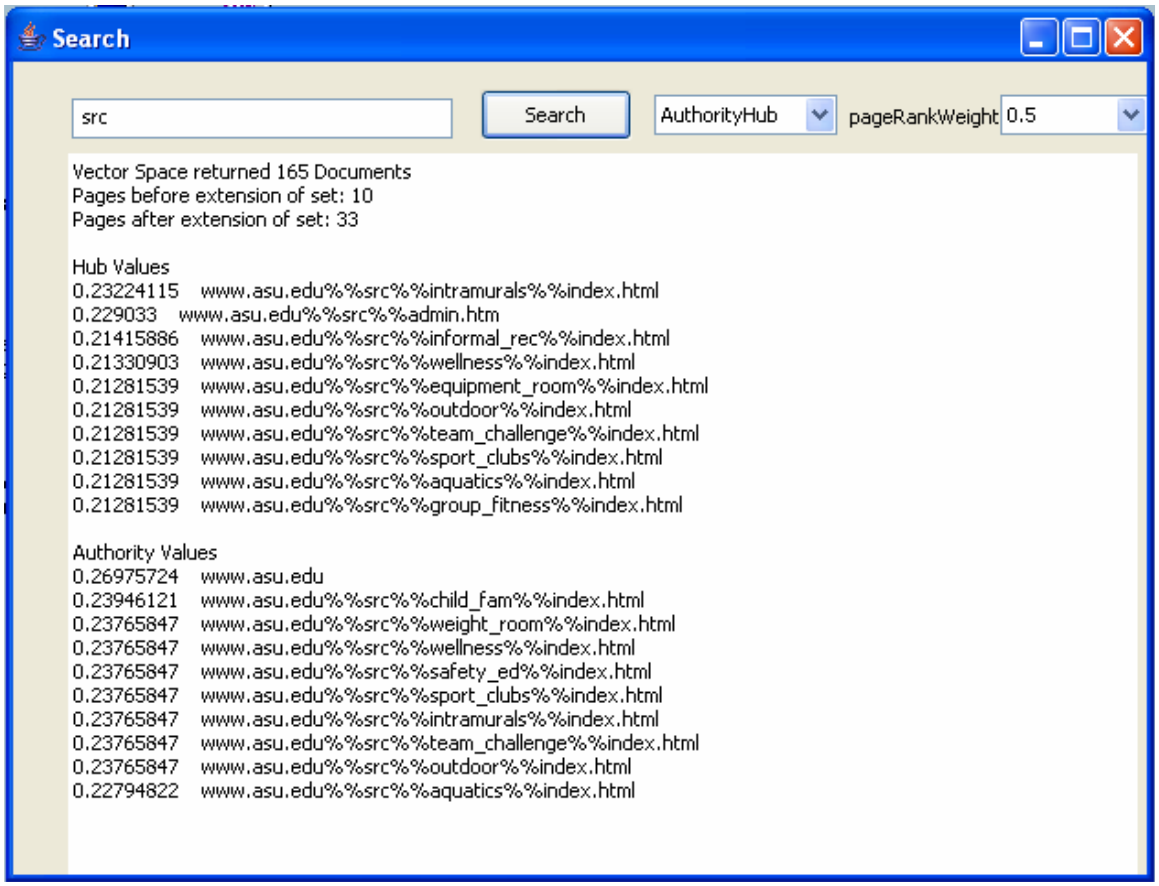
Value of C	Iterations to converge
0.8	14
0.6	8
0.4	5

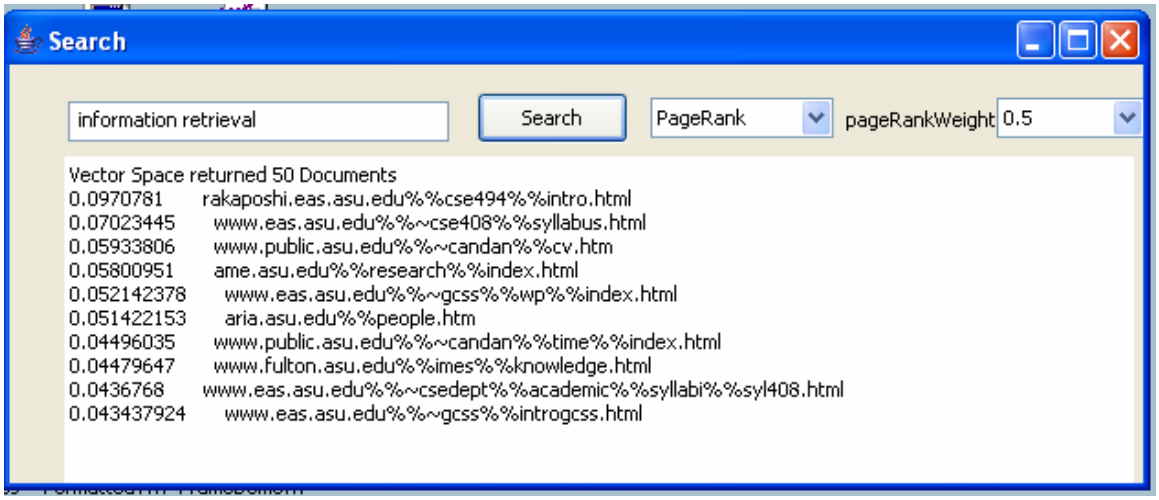
As the value of C decreases PageRank depends more on Random behavior. So PageRank value of each page depend less on links. And random behavior part is more stable. So PageRank values change less on each iteration. And convergence is guaranteed in a few iterations for small value of C.

Project B Report (GUI)

The GUI provides facility to input query, select ranking method (i.e., Vector Space, Authority Hub, PageRank) and also PageRank weight. Following are some snapshots of GUI.






Search
information retrieval
Search
PageRank
pageRankWeight
0.5
Vector Space returned 50 Documents
0.0970781 rakaposhi.eas.asu.edu%%cse494%%intro.html
0.07023445 www.eas.asu.edu%%~cse408%%syllabus.html
0.05933806 www.public.asu.edu%%~candan%%cv.htm
0.05800951 ame.asu.edu%%research%%index.html
0.052142378 www.eas.asu.edu%%~gcss%%wp%%index.html
0.051422153 aria.asu.edu%%people.htm
0.04496035 www.public.asu.edu%%~candan%%time%%index.html
0.04479647 www.fulton.asu.edu%%imes%%knowledge.html
0.0436768 www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl408.html
0.043437924 www.eas.asu.edu%%~gcss%%introgcss.html