MicroGoogle Search Engine

http://moselle.eas.asu.edu:8080/cse494/

Authority Hub Computation

Introduction

For a given query first we find the root set of the query using the Vector space model which uses the tf-idf weighting scheme as described in Proj A. We extract top K results using the vector space model for a given query terms. After extracting the root set, for each page p in the root set we find the set of pages that have links from the page p and add them to root set to get the base set. Also for each page p in the root set, we add the set of pages which point to p to the base set. Base set initially contains all the pages in the root set. This is how we generate the base set from the root set.

Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

Algorithm for Computing AH

Algorithm: **computeAH**
Input: Query, N
Output:Top N Authority and Hub pages

```
rootset = getTopKPagesVectorSpace(query,k)
baseset = null;
for each page p in rootset
        add p to baseset
        for each page pTo that p has a link to
                add pTo to baseset
        for each page pFrom that has a link to p
                add pFrom to baseset
end for
for each page p in baseset
        auth[p] = 1;
        hub[p] = 1;
end for
for iter = 1 to MAX_ITERATIONS
        for each page p in baseset
                for each q in baseset
                        if link(q,p)
                                auth[p] += hub[q];
                        end if
                end for
        end for
        for each page p in baseset
                for each q in baseset
                        if link(p,q)
                                hub[p] += auth[q];
```

```
                        end if
                    end for
                end for
                normalize authority and hub score for each page
                if ( converged )
                    break;
        end for

        for  i =0 to N
                display(auth[i]);
        for  i =0 to N
                display(hub[i]);
```

end **ComputeAH**

Convergence

If the sum of square of differences between the current authority values and the previous authority values is less than a certain predetermined threshold δ and also the sum of square of differences between the current hub values and the previous hub values is less than a certain predetermined threshold δ, then we say the authority and hub values have stabilized. Currently the threshold δ is kept at 0.0000001.

Time Complexity

Let,

K = size of the root set
T = Average number of pages that a page points to
F = Average number of pages that points a page
I = maximum number of iterations to converge
B = Average number of pages in the base set

$O(K*(T+F) + B + I *(B*B + B*B + B + B) + N + N) = O(I*B^2)$ since K,T,F are much smaller than B.

Space Complexity

We use a matrix to store the link structure. Since the size of base set is limited, we would not use too much memory in storing the link structure in matrix form of size B * B. We require two arrays of size B for storing current and previous authority values to check for convergence and similarly for hubs.

Comparison of results of A/H with those given by Vector Space Ranking

Good Authorities intuitively corresponds to pages which have more in-links and good hubs corresponds to pages which have more out-links. Due to nature of computation of authorities and hubs(based on link structure), it is possible that certain pages even though are not relevant to the query always show up in the top authorities and top hubs list. These are the pages which are pointed to by many different pages like www.asu.edu so they are always get added to base set as it is highly probable that one of the pages in the root set contains a link to these pages.

Solution
Use link context while giving weights to links. We could find the vector similarity between query terms and vicinity of characters (+-n) surrounding the anchor text and use that similarity as the weight of the link between two pages. This would eliminate pages which are pointed by many pages like www.asu.edu but are not relevant to the query from coming at the top of authorities and hubs list.

Consider the query: **software engineering**
Pages in the base set:10
Pages after extending the base set:84

Top 10 authorities

1. www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml 0.5131803846742273
2. www.asu.edu 0.1458154055171595
3. cpd.asu.edu%% 0.1289585507846079
4. www.eas.asu.edu%%CEAS%%students%%internships.shtml 0.1289585507846079
5. www.eas.asu.edu%%~industry%% 0.1289585507846079
6. www.eas.asu.edu%%CEAS%%students%%scholarships.shtml 0.1289585507846079
7. www.eas.asu.edu%%~csedept%% 0.1289585507846079
8. www.eas.asu.edu%%~mae%% 0.1289585507846079
9. www.eas.asu.edu%%CEAS%%students%%studentorg.shtml 0.1289585507846079
10. www.eas.asu.edu%%CEAS%%resources%%index.shtml 0.1289585507846079

Top 10 hubs

1. www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml 0.8945253444375031
2. www.eas.asu.edu%%CEAS%%resources%%index.shtml 0.0739905005345562
3. www.eas.asu.edu%%CEAS%%resources%%asures.shtml 0.0739905005345562
4. www.eas.asu.edu%%CEAS%%students%%studentorg.shtml 0.0739905005345562
5. www.eas.asu.edu%%CEAS%%students%%scholarships.shtml 0.0739905005345562
6. www.eas.asu.edu%%CEAS%%students%%collegeorg.shtml 0.0739905005345562
7. www.eas.asu.edu%%CEAS%%alumni%%profile.shtml 0.0739905005345562
8. www.eas.asu.edu%%CEAS%%resources%%ceasres.shtml 0.0739905005345562
9. www.eas.asu.edu%%CEAS%%research%%index.shtml 0.0739905005345562
10. www.eas.asu.edu%%CEAS%%depts%%index.shtml 0.0739905005345562

Here none of the pages in the top 10 authorities and top 10 hubs are relevant to the query, however they have high authority and hub scores as they are pointed to by many pages. Also there is an issue of ties here. Many pages seem to have the exact same authority score as other pages and also exact same hub score as other pages; hence it may be possible that we might miss a relevant result while displaying the just top 10 authorities and hub.

For the **Vector Space Model**, the results are:
1. www.eas.asu.edu%%~cse461%%top.html Score:0.5582
2. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl360.html Score:0.5502
3. www.eas.asu.edu%%CEAS%%depts%%degreeprograms.shtml Score:0.4955
4. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl462.html Score:0.4679
5. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl461.html Score:0.4529

6. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl562.html Score:0.4257
7. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl460.html Score:0.4206
8. www.asu.edu%%aad%%catalogs%%general%%engineering.html Score:0.3853
9. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl564.html Score:0.3727
10. www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl563.html Score:0.3564

The results given by the vector space ranking contain many relevant pages:
Pages 1,2,4,5,6,9,10 are relevant to the query.

Similarly for the other queries, we see that results given by vector space ranking are more relevant to the query. This is due to the fact vector space model ensures that pages returned contain at least some terms in the query which is not guaranteed for Authority/Hubs computation as it ranks the results on the importance based on the link structure after the root set is extracted.

Authorities vs Hubs

From the precision study, it is seems that authorities in general provide more relevant results than hub values. This seems to be intuitive as a page which is pointed to by many pages is more important than a page which points to many pages. We see that for all the test queries, authorities have same or higher precision than the hub values(see Figure 6 later when compared with PageRank).

Effect of changing the root set size on the size of the base set

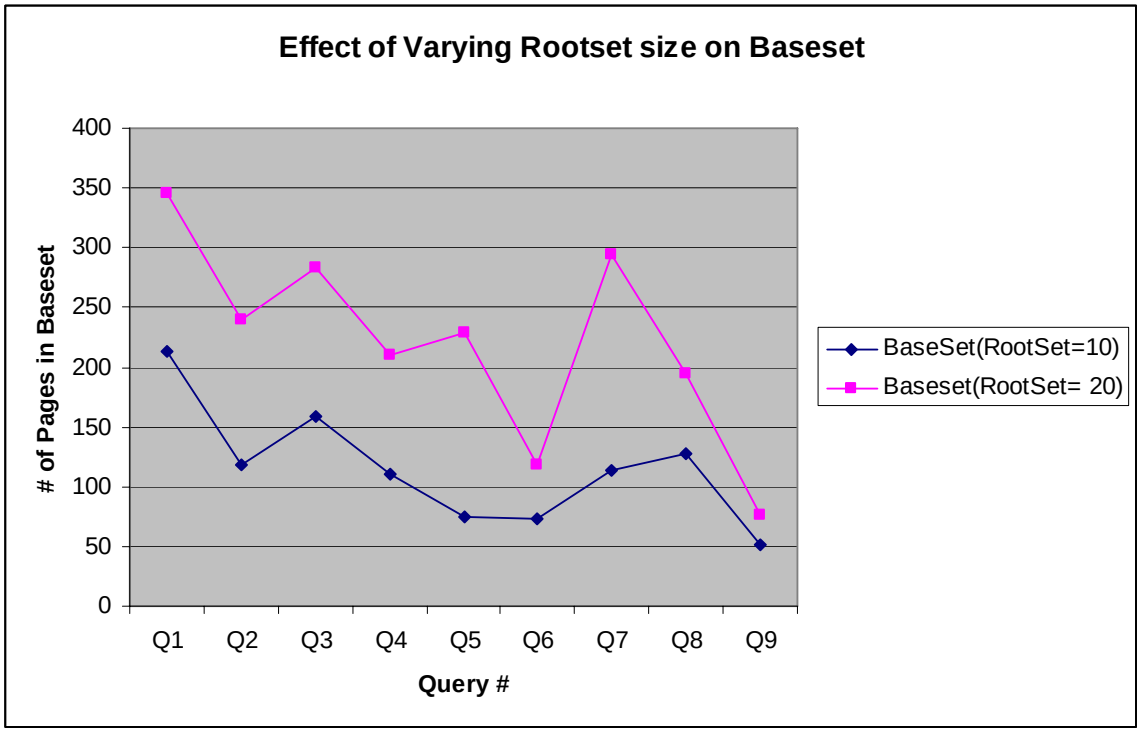


Figure 1: Effect of changing the root set size on the size of base set
This shows that as the size of the root set is doubled, the size of base set also increases by the same amount.

In order to see the effect of increase in the root size on the convergence rate, we calculated the number of iterations it takes for the AH values to stabilize.

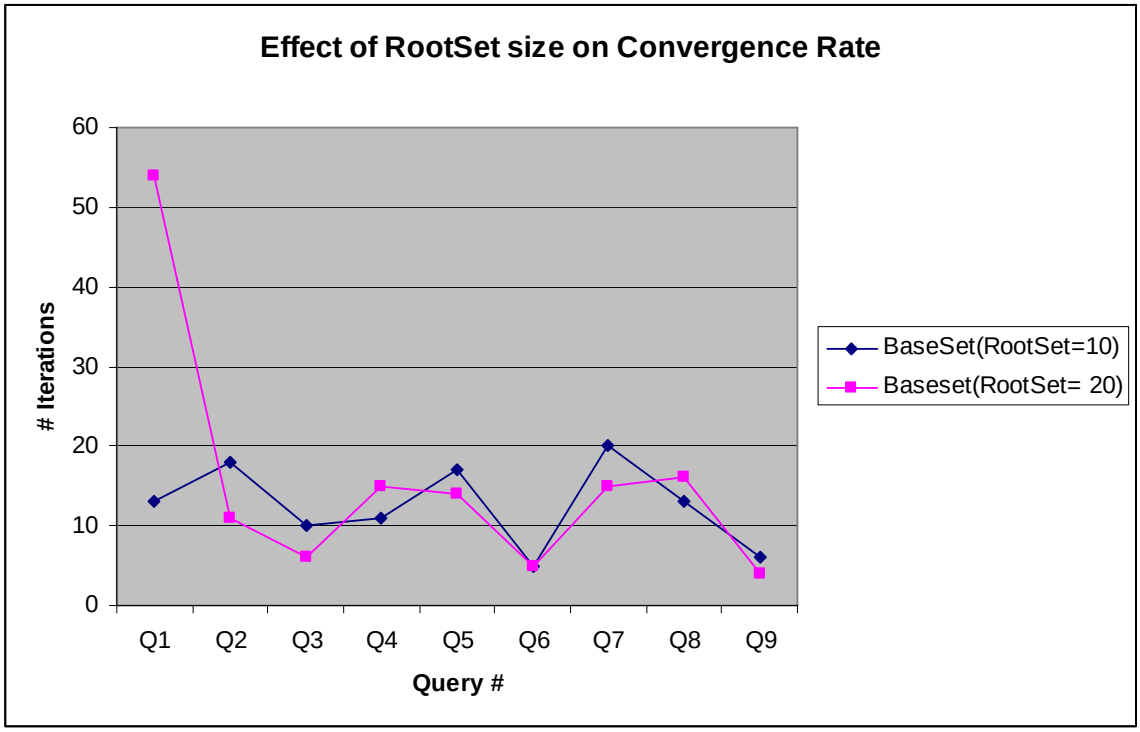


Figure 2: Effect of doubling the root set size on the convergence rate

This result seems to be some what counter-intuitive. As the size of the base set increases, we expect it to take longer time to converge (i.e. more iterations). However, for the given queries, iterations taken after doubling the root set size were almost same as those with root set size = 10.

Effect of change in root set size on results

Consider the query:Computer Science

For root set size = 10, the results are:

Searching for: computer science
Number of Docs in Index :7513
Query word:computer Total docs : 1357 Total Hits:1357
Query word:science Total docs : 2141 Total Hits:2724
Pages in the base set:10
Pages after extending the base set:75

Top 10 authorities

1. prism.asu.edu%%education_studentresearch.asp 0.70313592589279
2. www.asu.edu 0.1644901371959855
3. prism.asu.edu%%education_studentresearch.asp 0.1509195454260912
4. prism.asu.edu%%education_courses.asp 0.1509195454260912
5. prism.asu.edu%%resources_login.asp 0.1509195454260912
6. prism.asu.edu%%resources_links.asp 0.1509195454260912
7. prism.asu.edu%%education_mission.asp 0.1509195454260912
8. prism.asu.edu%%research_projects.asp 0.1509195454260912
9. prism.asu.edu%%research_publications.asp 0.1509195454260912
10. prism.asu.edu%%research_projectarchives.asp 0.1509195454260912

Top 10 hubs
1. prism.asu.edu%%education_studentresearch.asp 0.7806010878465711
2. prism.asu.edu%%about_mission.asp 0.13596167944172036
3. prism.asu.edu%%resources_links.asp 0.13596167944172036
4. prism.asu.edu 0.13596167944172036
5. prism.asu.edu%%research_3dk-pa.asp 0.13596167944172036
6. prism.asu.edu%%news_news.asp 0.13596167944172036
7. prism.asu.edu%%research_publications.asp 0.13596167944172036
8. prism.asu.edu%%education_mission.asp 0.13596167944172036
9. prism.asu.edu%%research_partnerships.asp 0.13596167944172036
10. prism.asu.edu%%education_courses.asp 0.13596167944172036

For root set size = 20, the results are:
Searching for: computer science
Number of Docs in Index :7513
Query word:computer Total docs : 1357 Total Hits:1357
Query word:science Total docs : 2141 Total Hits:2724
Pages in the base set:20
Pages after extending the base set:229
Top 10 authorities
1. www.eas.asu.edu%%~csedept%%Students%%StudentOrgs%%org.shtml 0.91956659236573
2. www.asu.edu 0.08271140652524597
3. cse.asu.edu 0.0712607792187009
4. www.asu.edu 0.06844406126328927
5. www.eas.asu.edu%%~csedept%%research%%researchCenters.shtml 0.064178186646883
6. www.eas.asu.edu%%~csedept%%research%%researchInstitutes.shtml 0.0641781866468837
7. www.eas.asu.edu%%~csedept%%research%%affResearchCenters.shtml 0.06417818664688
8. www.eas.asu.edu%%~csedept%%research%%researchAreas.shtml 0.06417818664688374
9. www.eas.asu.edu%%~csedept%%research%%research.shtml 0.06417818664688374
10. www.eas.asu.edu%%~csedept%%people%%staff%%staff.shtml 0.06417818664688374
Top 10 hubs
1. www.eas.asu.edu%%~csedept%%Students%%StudentOrgs%%org.shtml 0.44838394795029
2. www.eas.asu.edu%%~csedept%%AcademicPrograms%%Undergraduate%% 0.15449805640
3. cse.asu.edu%%AcademicPrograms%%Undergraduate%% 0.15449805640626824
4. www.eas.asu.edu%%~csedept%%Students%%Scholarships%%scholarships.shtml 0.1344240
5. www.eas.asu.edu%%~csedept%%people%%QuickContactInfo.shtml 0.13166789932370368
6. www.eas.asu.edu%%%7Ecsedept%% 0.13166789932370368
7. www.eas.asu.edu%%~csedept%%people%%faculty%%faculty.shtml 0.13166789932370368
8.www.eas.asu.edu%%~csedept%%news%%Announcements%%Defenses%%Summer2003Defenses.shtml 0.1316678993
9. cse.asu.edu%%AcademicPrograms%%Graduate%% 0.13166789932370368
10. cse.asu.edu%% 0.13166789932370368

The top 10 authorities and hub pages for root set size = 20 are completely different from that obtained with root set size = 10. Pages obtained from root set size = 20 are more relevant than that obtained from root set size = 10. This shows that changing the root set size can affect the precision of the search results.

Effect of change in root set size on Search Time

However the time taken for getting the search results was almost doubled for root set = 20 because we need to retrieve more pointed to pages and more pointed by pages corresponding to the pages in the root set and this is done online after the query is issued.

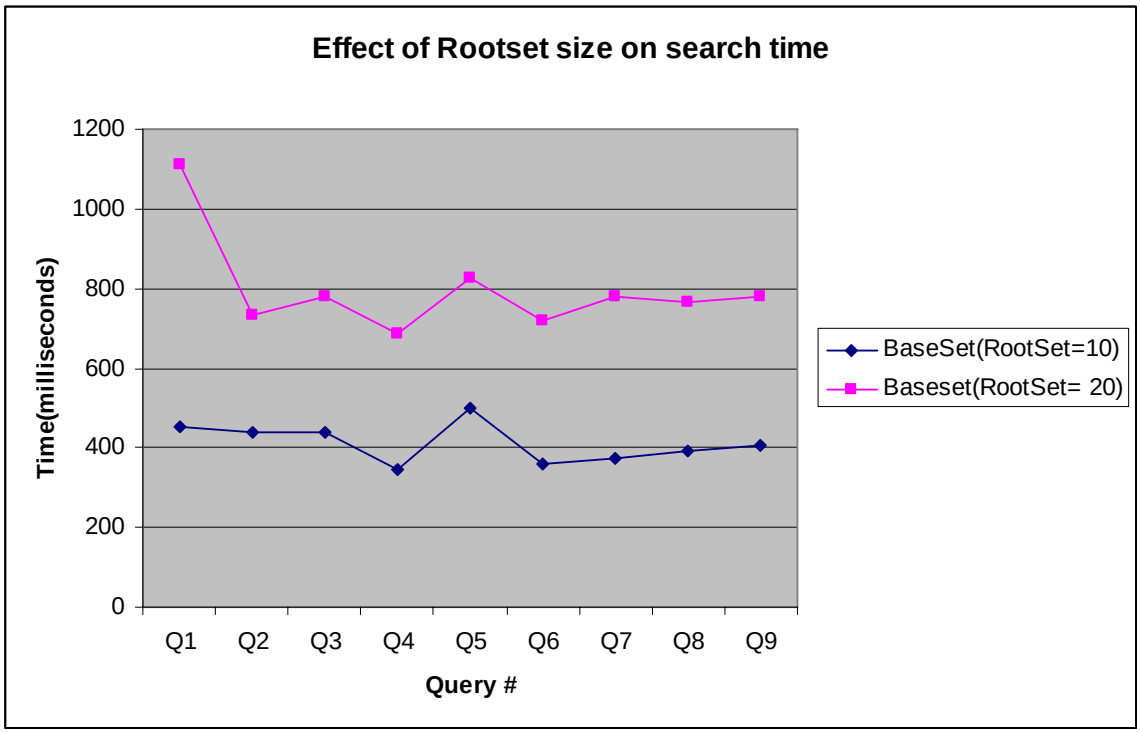


Figure 3: Effect of increase in the root set size on the query execution time

Tyranny of Majority

For the query "Fall Semester", we observed the phenomenon of tyranny of majority.

Top 10 authorities

1. www.asu.edu%%registrar%%general%%semcal.html	0.9999999481027821
2. geology.asu.edu%%courses%%index.html	1.25760162148935E-4
3. www.west.asu.edu%%	1.0331434675628713E-4
4. www.east.asu.edu%%	9.025821417670335E-5
5. www.asu.edu	7.756407919955823E-5
6. www.asu.edu%%xed%%	6.672793996281735E-5
7. www.asu.edu%%	6.672793996281735E-5
8. www.asu.edu%%copyright	5.278216070858136E-5
9. www.west.asu.edu%%asadv%%	4.2308019074279146E-5
10. www.west.asu.edu%%search%%index.htm	4.2308019074279146E-5

Top 10 hubs

1. www.west.asu.edu%%sa%%registrar%%index.htm	0.13867504487399762
2. asuonline.asu.edu%%registration%%index.cfm	0.13867504487399762
3. www.west.asu.edu%%chs%%sw%%	0.13867504487399762
4. www.east.asu.edu%%ecollege%%wellness%%html%%calendar.htm	0.13867504487399762
5. www.west.asu.edu%%chs%%rtm%%	0.13867504487399762
6. www.west.asu.edu%%chs%%RTM%%index.htm	0.13867504487399762

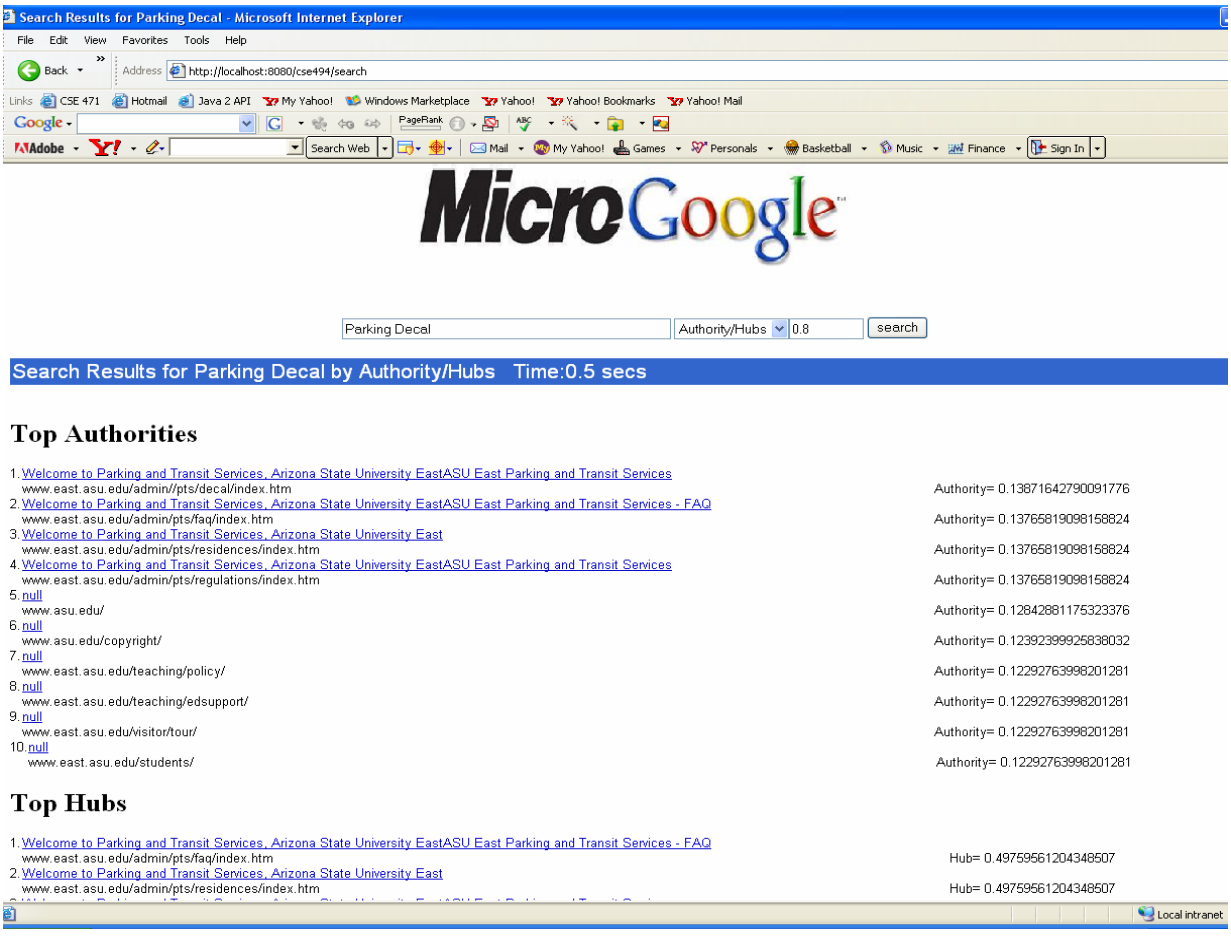
7. www.west.asu.edu%%chs%%sw%%faculty.htm 0.13867504487399762
8. www.east.asu.edu%%ctas%%ecet%%Announce%%announces.html 0.13867504487399762
9. www.east.asu.edu%%ctas%%mmet%%Announce%%announces.html 0.13867504487399762
10. www.asu.edu%%calendar 0.13867504487399762

Here, the authority value of the page "www.asu.edu%%registrar%%general%%semcal.html" is close to 1 while authority values of all other pages in the top 10 are close to 0. This might be because there are separate disconnected communities and the page which has the highest authority belongs to the community where there are more pages pointing to the above page than in any other community.

This phenomenon was not observed in the hub values as it appears that all the top 10 hub pages have almost the same hub values.

GUI

A servlet based GUI is developed to present results to the users in a Googlish style.



It shows the title, url and the authority score of the top 10 authorities and title, url and the hub score of the top 10 hubs as shown in the figure above. It also shows the time taken to search.

PageRank

Since we are dealing with large number of pages of the order of 10000, the algorithm for computation of pagerank has to be fast with efficient use of memory. We do not construct the transition matrix M* explicitly since it would take a lot of space as each element of M* is of the type double. Rather, we make use of the fact that original adjacency matrix would be sparse which is typical for the web graph. Hence we use adjacency list representation to store out-links for each page. This leads to a tremendous reduction in the amount of memory used as well as speeds up the page rank computation.

Algorithm

Returning search results using the weighted combination of page rank and vector space similarity is divided in two parts:

Offline Computation:

To get results using the Vector Space Model, we first compute the norm of each document offline based on the tf-idf weighting scheme and store the results in a hashtable for efficient retrieval. We also store (url,title) pair in another hashtable to display the title for each url in the search result. Computation of page ranks for all the crawled pages in the repository is done only once when the server is started. www.asu.edu turns out to be the page with highest page rank.

Algorithm ComputePageRank;
Input: crawled pages directory, link structure file
Output: hashtable containing the urls and their page rank

```
N = total crawled pages in the link structure file
pageList = list of all crawled pages in the link structure file
for each crawled page p in link structure file
        adjList[p] = get all pages that p points to;
        pagerank[p] = 1/N
end for
for iter = 0 to MAX_ITERATIONS
        for each page p in PageList
                if p is not a sink node
                        for each page q ϵ adjList[p]
                                pagerank[q] + = pagerank[p] / adjList[p].size
                        end for
                else
                        pagerank[q] + = pagerank[p] / N
                end if

        end for
        for each page p in PageList
                pagerank[p]  = DAMPING_FACTOR * pagerank[p] +
                                        (1-DAMPING_FACTOR)/N
        end for
        if  converged
                break;
```

```
        end for
        normalize by dividing each page rank with maximum value of page rank
        for each page p in PageList
                add (url[p],pagerank[p]) to pageRankHashTable
        end for
        return pageRankHashTable
end ComputePageRank
```

Convergence

If the sum of square of differences between the current page rank values and the previous page rank values is less than a certain predetermined threshold δ, then we say the page ranks have stabilized. Currently the threshold δ is kept at 0.0000001.

Time Complexity

Let N = total number of crawled pages in the repository
I = max number of iterations to converge
P = average number of pages pointed to by a page
O(N + I * (N * P + N + N) + N + N)= O(I*N*P)

Space Complexity

We need an adjacency list array for storing the adjacency list for each page. So total memory require is N * P instead of N * N if we were to use transition matrix M* as P <<< N. Then we require two more arrays of size N to store the current ranks and previous ranks to check for convergence.

Online Computation

The online part consists of retrieving the vector space search results for the query and combining the vector space similarity score with the page rank computed offline to get a combined weighted score. The search results are then displayed in ascending order of this combined score.

Algorithm PageRankVectorSpace
Input: query,Weight W
Output: ranked list of results

```
        pagesVS = getAllPagesVectorSpace(query)
        for each page p in pagesVS
                pagerank = getPageRank(p);
                combinedScore = W * pagerank + (1-W) * VectorSimilarity(p);
        end for
        sort(combinedScore)
        display top N pages with highest combined score
end PageRankVectorSpace
```

Time Complexity

Let,
N= average number of pages returned by the Vector Space Model for given query

Tn be the number of terms in the query Q

H be the number of hit documents for the query Q

Time complexity of retrieving pages from the vector space model O(Tn * H)

Time complexity of the combined method is O(Tn * H + N + N log N) = O(N log N+Tn*H)

PageRank Computation

Time Taken(sec)	Total Pages	Highest Page Rank	Damping Factor	# of Iterations to Converge	Threshold
24.39	9972	www.asu.edu	0.9	8	0.0000001
24.67	9972	www.asu.edu	0.8	6	0.0000001
23.93	9972	www.asu.edu	0.7	4	0.0000001
23.93	9972	www.asu.edu	0.6	4	0.0000001
23.89	9972	www.asu.edu	0.5	3	0.0000001
23.79	9972	www.asu.edu	0.4	3	0.0000001
23.76	9972	www.asu.edu	0.3	3	0.0000001
23.67	9972	www.asu.edu	0.2	2	0.0000001
23.5	9972	www.asu.edu	0.1	1	0.0000001

Effect of changing Damping Factor

Varying the damping factor between 0.7 to 0.9 does not have a major change on the order of results. However, for significantly low values say [0.1.0.3], probability of the random surfer moving to any random page is higher than following the link structure and hence there is some change in the results.

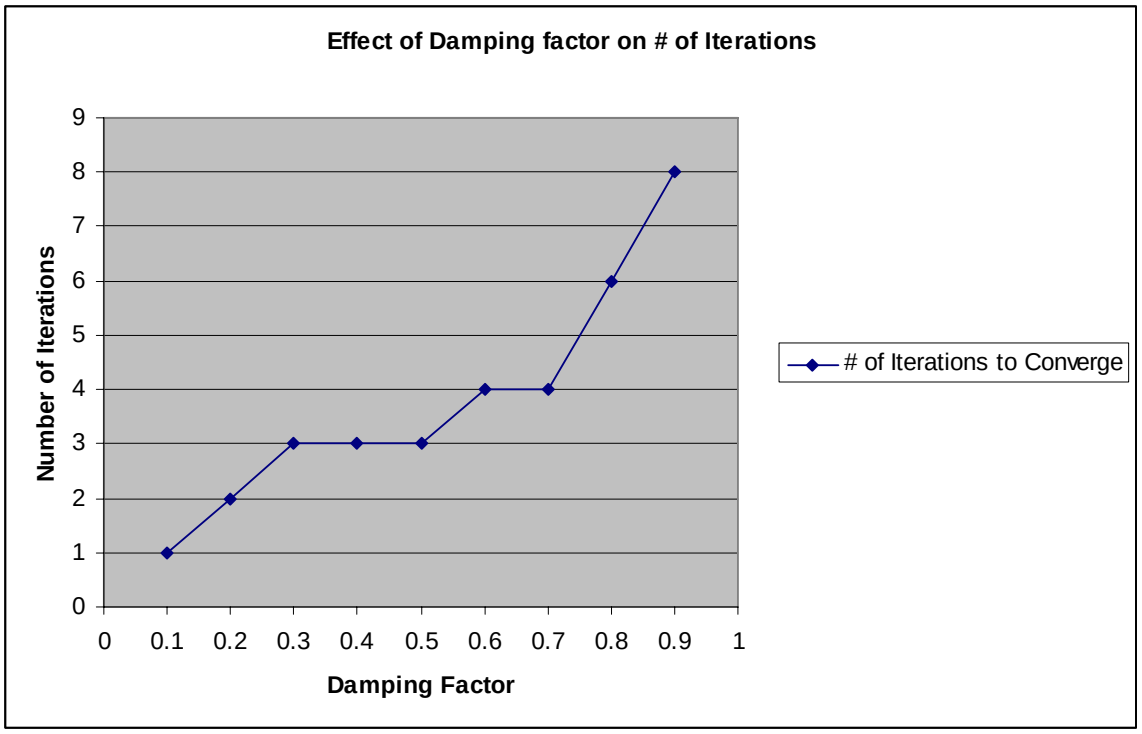


Figure 4: As the damping factor corresponding to the random surfer model increases, the number of iterations to converge also increases.

We then normalize the page rank by dividing each the page rank of each page by the maximum page rank, so that for each page the page rank is between [0.1].

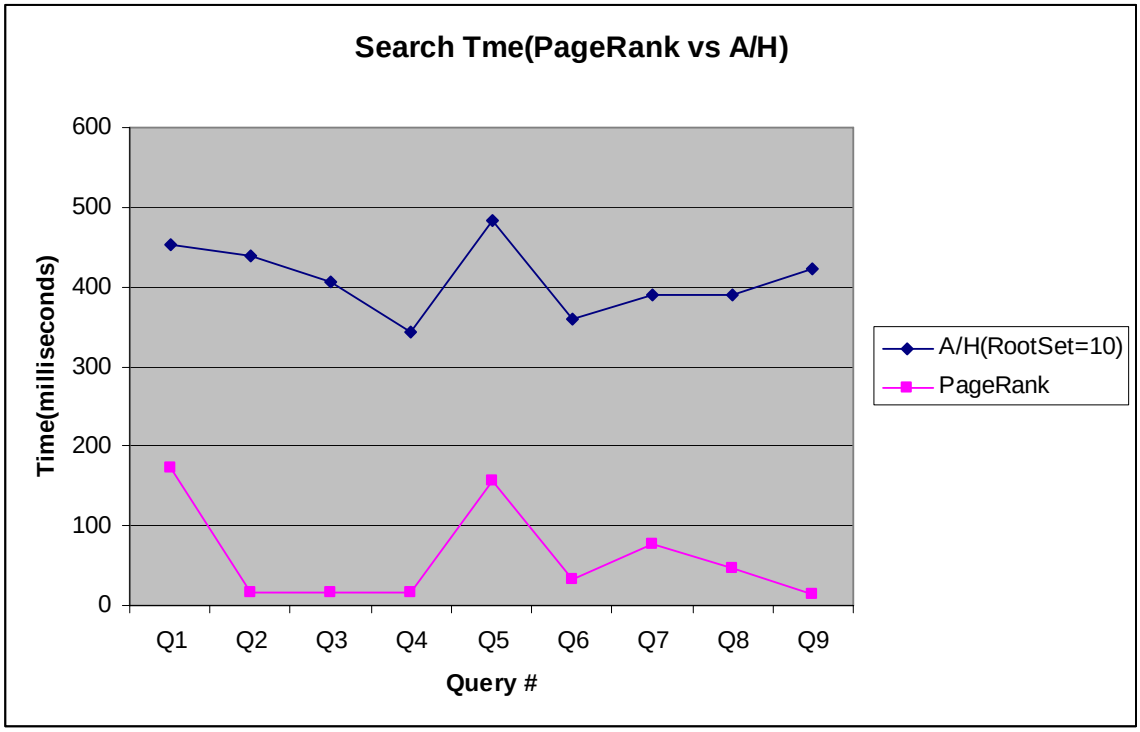


Figure 5: Comparison of search times for authority/hubs and page rank

This shows that Authority/Hubs take almost twice as much time as that taken by Page Rank. This is due to the fact that all the processing for computing the forward set and backward set to construct the base set is computed online after the query is issued, whereas the page rank for each page is computed offline and retrieved quickly after the query is issued.

Precision Study[1]

Query	Authority(Root size 10)	Hub(Root size 10)	PageRank+Vector Similarity
Fall Semester	0.2	0.2	0.5
Information Retrieval	0.3	0.3	0.8
Transcripts	0.2	0.2	0.4
Networks	0.2	0.2	0.7
Computer Science	0	0	0.5
Multimedia Database	0.3	0.2	0.4
Software Engineering	0	0	0.7
Parking Decal	0.4	0.4	0.9
SRC	0.4	0.6	0.7

This shows that PageRank + Vector similarity always give better precision than authorities and hubs. This is due to the fact that page rank considers results from the

[1] For each query, the top 10 results were manually judged as either relevant or not relevant and the precision was computed based on that.

vector space results as well as combining importance of the page derived from page rank. Authority/hub considers only link structure once the base set is extracted, hence more popular but irrelevant pages appear at the top.

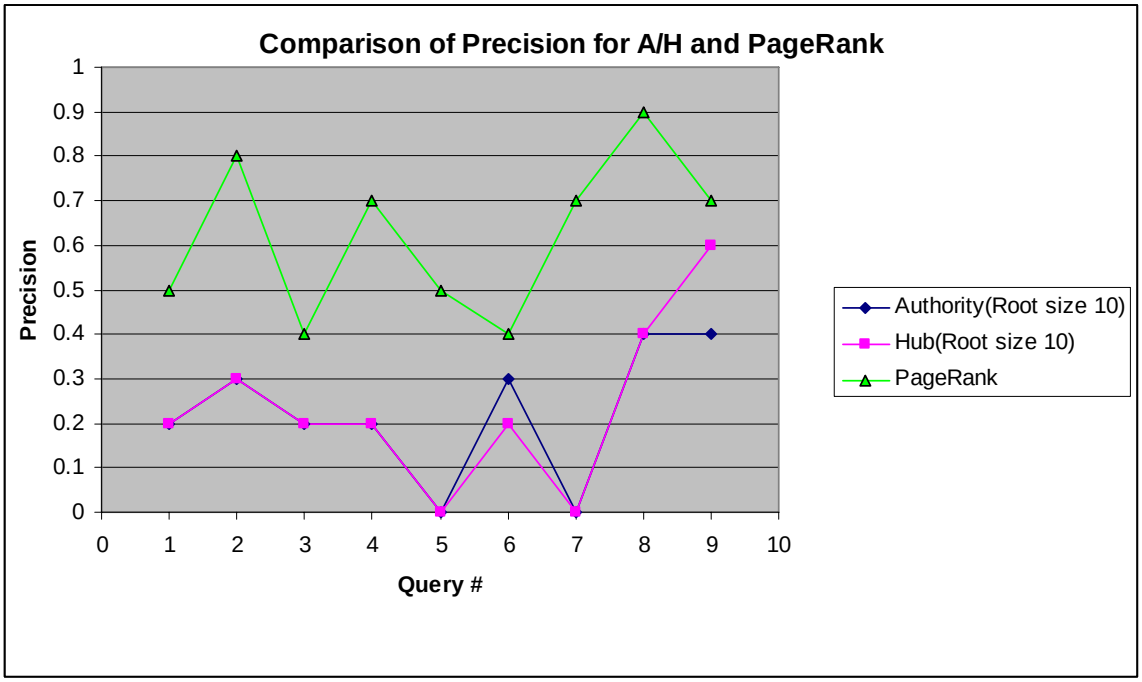


Figure 6: Comparison of precision for authority/hubs and page rank

Handling Dangling Links and Sink Nodes

We do not consider dangling links for computation of the page rank. They are removed while building the adjacency list for each crawled page.

In order to take account for the sink nodes, we consider the random surfer model where the user can go from the sink nodes to any other page with equal probability. Hence the page rank of the sink nodes gets divided equally among all the other pages.

Effect of Normalization Page Rank

The relative order of the results is not affected if page ranks are not normalized in the range [0,1]. Only the combined similarity score(Vector Space + Page rank) will be different.

Results Comparison with current Google Top Pages

Query	Search Engine
Transcripts	MicroGoogle(Page Rank) Top 3 1. www.asu.edu%%admissions%%steps%%finaltranscripts.html 2. www.asu.edu%%registrar%%transcripts%%index.html 3. www.asu.edu%%admissions%%steps%%transcripts.html
	Google Top 3

	www.asu.edu/registrar/transcripts/ www.asu.edu/admissions/steps/transcripts.html www.kaet.asu.edu/horizonte/transcripts.html
Information Retrieval	MicroGoogle(Page Rank) Top Page: rakaposhi.eas.asu.edu%%cse494%%intro.html
	Google Top Page: rakaposhi.eas.asu.edu/cse494

Although the crawl is quite old, the top results of Google Search Engine with MicroGoogle for some queries winds up being the same.

Effect of Weight W given to PageRank

We normalize the page ranks so that each page has page rank between [0,1]. However, due to the large size of the courpus, it turns out that most of the pages have very low page rank in the order of 10^{-2}. Hence when page rank score is combined with vector similarity which has values of the order of 10-1 we need to have a weight which does give significant weight to the page rank in order to make an effect on the rankings otherwise the results obtained by fall semester would be same as those obtained by the Vector Space Model. This happens when the value of weight given to page rank is 0.5 and also 0.5 to the Vector Space Model. Thus effect of page rank is unnoticeable in the results. To overcome this, we give a very high weight to page rank 0.9. This results in giving more relevant results:

For the given query: Fall Semester

(Weight given to PageRank is 0.5)
(Weight given to Vector Space Similarity is 0.5)
1. www.asu.edu%%registrar%%general%%semcal.html 0.2108526665270732
2. www.eas.asu.edu%%~cse355%%cse355B%%index.html 0.1954478306730186
3. coe.asu.edu%%oss%%prereg%%index.php 0.1893845628579164
4. geology.asu.edu%%courses%%index.html 0.1497374687229725
5. www.eas.asu.edu%%~cse430%%logo_page.html 0.13726308347432578
6. www.eas.asu.edu%%~cse512%%home512.html 0.12597222504907746
7. www.west.asu.edu%%amerstud%%Degrees%%2003-2004%%Hisminor.htm 0.113462311677114
8. www.west.asu.edu%%ams%%Degrees%%2003_2004%%Hisminor.htm 0.11098476714851757
9. www.eas.asu.edu%%~hsarjou%%Courses%%coursesCSE.htm 0.10821807914100999
10. www.asu.edu%%admissions%%steps%%application.html 0.10752199108117919

Here only pages 1,4,5,6,9 are relevant to the query. Here the most relevant page: www.asu.edu%%calendar%%academic.html does not appear in top 10. Changing the weight value of Page Rank to 0.9 we get the most relevant page in the top 10 results.

(Weight given to PageRank is 0.9)
(Weight given to Vector Space Similarity is 0.09999999999999998)
1. www.asu.edu%%asunews%%index.html 0.05141964131573886
2. www.asu.edu%%registrar%%general%%semcal.html 0.04849039170700105
3. www.asu.edu%%asunews%%video.htm 0.04823336915999415
4. www.eas.asu.edu%%~cse355%%cse355B%%index.html 0.03976328172326558
5. coe.asu.edu%%oss%%prereg%%index.php 0.039381744378280965
6. www.asu.edu%%calendar%%academic.html 0.038382485918351195
7. geology.asu.edu%%courses%%index.html 0.03137951357757988

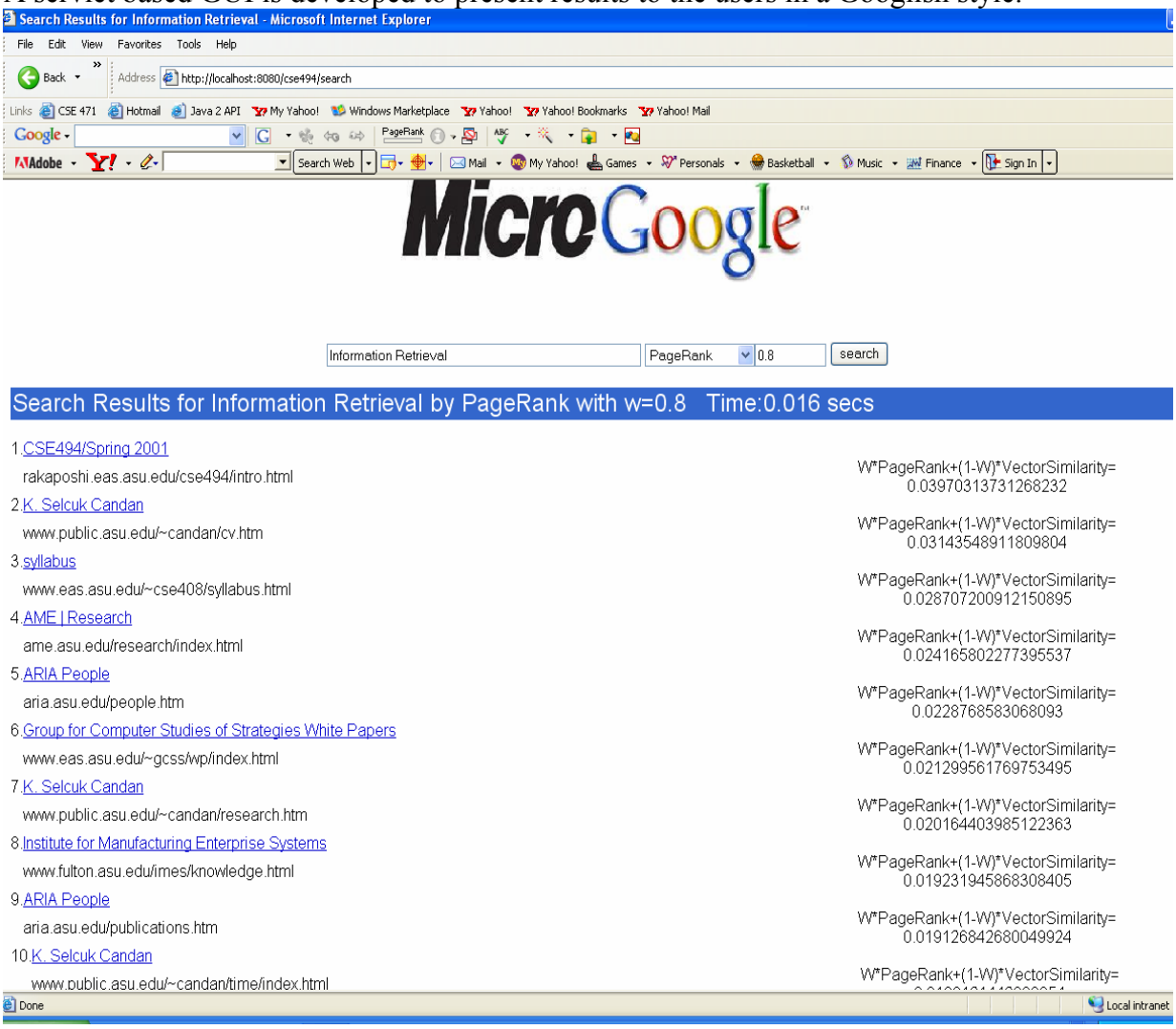
8. www.asu.edu%%registrar%%registration%%ways2reg.html 0.0306733359993054
9. www.asu.edu%%asunews%%arts%%arts_index.htm 0.028277416708143703
10. www.eas.asu.edu%%~cse430%%logo_page.html 0.02804805227310026

However for some of the queries, this had a deteriorating effect with most popular pages showing up in the top 10 even though they are not relevant to the query(for e.g. www.asu.edu/asunews/index.html).
Finding an appropriate value of W could be done by using **relevance feedback.**

If weight W given to Page Rank is very low, say 0.1, then the results of the Page Rank + Vector Space would be very similar to that of Vector Space Model as page rank is almost neglected in this case.

GUI

A servlet based GUI is developed to present results to the users in a Googlish style.



It shows the title, url and the combined weighted score(Vector Space + Page Rank) of the top 10 pages as shown in the figure above. It also shows the time taken to search.