CSE494 Information Retrieval

Project B Report 1 (A/H Computation)

Introduction

Project B task 1 implements the algorithms of Authority/Hub computation on the given crawled web pages in the asu domain. This report contains the following parts: First the algorithm and some implementation issues are discussed, followed by a comparison of performance of A/H ranking and the Vector Space Similarity ranking. Next the effects of varying the size of the root-set on the A/H computation and results are discussed. Then we will talk about some observations of the relevance of the pages returned by A/H computation. At last the efficiency issue will be analyzed briefly. The testing results of the given query are also attached.

Algorithm and Implementation

The basic idea of A/H computation is based on the facts that a page referenced by lot of important pages is more important (authority) and a page that references a lot of important pages is also important (hub). The algorithm can be briefly described as: for each given query, certain query method (for example, Vector Space Ranking) is applied on the available pages to retrieve a "root set" which contains the top pages considered most relevant by the query method. Then the root set is expanded by adding the pages pointing to and pointed by the pages in the root set. The expanded set is the base set for the A/H computation. Then for each page in the base set, it has an initial authority value and an initial hub value. The A/H computation is done iteratively. In each of the iterations, the new authority value of a given page is the sum of the hub values of all the pages pointing to it; and the new hub value of a given page is the sum of the authority values of all the pages it points to. This computation will eventually converge and at last the pages with the highest authority value will be shown to the user.

The implementation of this algorithm is straightforward: the given API provides the facilities to retrieve the pages pointing to and pointed by a given page. So in the implementation, for a given user query, we first retrieve all the relevant pages using the Vector Space Similarity algorithm implemented in project A, and take the top k pages as the root set. Then root set is expanded to base set by just simply going through every page in the root set and adding pages linking to and linking by it into the set. For the A/H computation, two hash tables are used to represent the authority/hub vectors respectively. The vectors have the page names as the keys and the authority/hub values as the values. In each iteration, first go through the authority vector and for each page, find the set of pages pointing to it and look up the corresponding hub values of these pages and add these values up to be the new authority value of that page; and then go through the hub vector and for each page, find the set of pages it points to and look up the corresponding authority values of these pages and add them up to be the new hub value of that page. After each iteration, the authority and hub vectors need to be normalized by dividing all the values by the lengths of the vectors respectively so that the values are always between 0 and 1. To judge the convergence, before each iteration a copy of the authority vector is kept and after the iteration, the difference of the new vector and the old one are calculated. This difference is compared to a given threshold and if it is below the threshold we consider the computation converged. In the implementation the threshold of 0.01 is used and it seems to work good enough. After the A/H computation, the pages with the top 10 authority values and top 10 hub values are shown to the user.

Comparison of Performance of A/H Ranking and Vector Space Similarity Ranking

The Vector Space Similarity algorithm ranks the pages totally based on the similarity of the documents and the given query in the vector space of all the terms, regardless the “importance” of the pages. A/H ranking take the importance of pages into consideration and are more likely to show some “important” pages to the user.

The testing on the given data set shows that the A/H computation will find the pages which are important in terms of number of references. If the user is interested in finding “important” pages about some given topic A/H ranking will help because in the pure Vector Space Similarity model, a page that is not important but contains exactly the same content with the query might be returned to the user as the most relevant one.

But an important observation is that although the A/H ranking will find more “important” pages, these pages are not necessarily highly relevant to the query. The reason is that the way “root set” is expanded does not take the relevance issue into consideration, so for some very popularly cited pages, for example the page www.asu.edu in the testing setting, many pages have a link to it and thus it is almost always added into the base set and has the highest authority value although it is not really relevant to the query. A possible way to handle this problem is, during the root set expansion, only add the relevant pages into the base set so that the final top A/H pages are still relevant.

The Effect of Varying the Size of “Root Set”

In the testing three values of root set (10, 15 and 20) is used to compare the effect of the size. The first observation is the that when the size of the root set increases, the expanded base set increases the size almost proportionally which reflects the average numbers of forward/back links of the pages in the collection.

When the size of the root set is changed, we can find that the final top 10 authority/hub pages changed dramatically (except for the top authority pages, which are really popular and are less likely to be effected by the topology of the link graph of the base set). For example, when the size is changed from 10 to 20, the top 10 hub pages of the query “Multimedia Database” changed from

```
Top 10 Hub pages:
Rank     Hub Value                URL
-------- ----------------------   ------------------------------
0        0.2603070609106187       www.asu.edu%%index
```

```
1       0.2603070609106187     www.asu.edu%%index%%
2       0.2603070609106187     www.asu.edu%%index%%index.html
3       0.25509813406743204    www.asu.edu%%programs%%index.html
4       0.25509813406743204    www.asu.edu%%programs
5       0.25509813406743204    www.asu.edu%%programs%%
6       0.25509813406743204    www.asu.edu%%asuweb%%entrance%%academic%%
7       0.22101238621464459    www.asu.edu%%aad%%catalogs%%general%%
8       0.21849883470765288    www.asu.edu%%aad%%catalogs%%
9       0.20200614179150264    www.asu.edu%%aad%%catalogs%%general%%index.html
```

to

```
Top 10 Hub pages:
Rank    Hub Value              URL
-----   ----------------       --------------------------
0       0.19087821898034235    www.asu.edu%%lib%%noble%%eng%%con101.htm
1       0.19076691055529799    www.asu.edu%%lib%%noble%%
2       0.19076691055529799    www.asu.edu%%lib%%noble
3       0.18863685257170562    www.asu.edu%%index
4       0.18863685257170562    www.asu.edu%%index%%
5       0.18863685257170562    www.asu.edu%%index%%index.html
6       0.1839526240869909     www.asu.edu%%lib%%noble%%chem%%
7       0.16674904666392198    www.asu.edu%%aad%%catalogs%%
8       0.16589465033735948    www.asu.edu%%programs%%index.html
9       0.16589465033735948    www.asu.edu%%programs
```

We find that for many queries, increasing the size of the root set seems to more or less find more “relevant” pages in the final output, because the root set contains only relevant pages and more relevant pages are involved in the A/H computation if root set is larger, so it is more likely to find some “relevant” pages with higher A/H value.

The reason of the dramatic effect of changing the size of the root set is that the A/H model totally makes the ranking based on the link structure. In another word, it is very sensitive to the topology of the base set link graph. And as we can see, increase the size of the root set will change the topology of the base set link graph a lot, so the A/H ranking will also change greatly. This also reinforces the point that the A/H ranking is not stable.

Relevance of Authority/Hub Pages

In the specification of A/H algorithms, the final output is the top k pages with the highest authority values. So an intuitively expectations of the final result is that the “authority” pages might be more relevant than the “hub” pages. But in the experiment setting of this project, it turns out that this is not true in most cases. Actually for the given queries, most of the “hub” pages seem to be more relevant than the “authority” pages. For example for the given query “parking decal”, the result is:

```
Number of Docs in Root Set : 10
Number of Docs in Base Set : 105

Top 10 authority pages:
Rank      Authority Value           URL
--------  ----------------------    ------------------------------
0         0.16270222708677431       www.asu.edu
1         0.1624753194663338        www.asu.edu%%
2         0.1576907596344321        www.asu.edu%%copyright%%
3         0.1554119992787901        www.east.asu.edu%%
4         0.15499600205434624       www.west.asu.edu%%
5         0.15459839524924004       www.asu.edu%%privacy%%
6         0.15305273355825483       www.asu.edu%%xed%%
7         0.1514303240601692        www.east.asu.edu%%admissions%%
8         0.15131063977154963       www.east.asu.edu%%about%%weather%%
9         0.15131063977154963       www.east.asu.edu%%about%%

Top 10 Hub pages:
Rank      Hub Value                 URL
--------  ----------------------    ------------------------------
0         0.19413580823094717       www.east.asu.edu%%about%%personnel%%
1         0.1533620259614583        www.east.asu.edu%%admin%%pts%%
2         0.1502286675951879        www.east.asu.edu%%contact%%
3         0.1499845536048614        www.east.asu.edu%%admin%%pts%%residences%%index.htm
4         0.1499845536048614        www.east.asu.edu%%admin%%pts%%parkingsafety%%index.htm
5         0.1499845536048614        www.east.asu.edu%%admin%%%%pts%%decal%%index.htm
6         0.1499845536048614        www.east.asu.edu%%admin%%pts%%events%%index.htm
7         0.1499845536048614        www.east.asu.edu%%admin%%pts%%faq%%index.htm
8         0.1499845536048614        www.east.asu.edu%%admin%%pts%%shuttle%%index.htm
9         0.1499845536048614        www.east.asu.edu%%admin%%pts%%appeals%%index.htm
```

In the above query results, apparently the hub pages are more relevant. The same thing happens to most of the given queries. Pages like www.asu.edu or www.asu.edu/copyright get very high authority value just because in this domain so many pages have links to them although they are not relevant to any given query.

The explanation of higher relevance of the hub pages might be, in most cases, the authority pages are those which serve as “roots” of subgraphs of the web and contains more “general” information (which are not likely to be very relevant to any specific topic), while the hub pages are more likely to be the “bottom” pages talking about some specific topics and having lots of links to the higher level pages. Thus for a given topic (query), the hub pages are more likely to be relevant.

Efficiency Issue

The A/H computation is done for every query and it turns out that although the number of the pages in the base set is not large (for a root set of size 10, the base set is usually around 100), the A/H iterations are still time consuming. It takes couple of seconds for a single iteration. In my implementation a single A/H iteration even takes longer time than a single pageRank iteration. The main reason is that the A/H computation needs to do the normalization in each iteration which involves lots of float number division and square root computation, which are largely avoided in the pageRank algorithms. So in the real system, if the number of queries in single time unit is large, the A/H computation might takes lot of computation resources.

CSE494 Information Retrieval

Project B Report 2 (pageRank)

Introduction

This part of project B implements the pageRank algorithm on the given crawled pages. This report contains the following parts: Fist the pageRank algorithm and the implementation are discussed followed by some important observation in the implementation. Then the result of A/H ranking and pageRank is compared and analyzed. Then we will talk about the effect of changing the weight of pageRank on the final ranking value, and the effect of varying the damping factor on the computation. Finally the efficiency issue is stated briefly. The testing results are also attached.

Algorithms and Implementation

PageRank use the link information to measure the global importance of the pages based on the backlinks (or citations) of the pages. The algorithm basically models the importance of a given page as the probability that the surfer finds himself on the page. If a page links to n pages, then the probability that the surfer follow any of the links is 1/n. That way the importance of this page will be propagated to the linked pages. At the same time, this model also considers the fact that at any time, the surfer might not follow any link on the given page, but start on another page randomly. So in this model there would be a small probability that the surfer goes from one page to another even if they are not linked.

The basic algorithm is, first initiate the pageRank of all pages to be 1/N, where N is the number of the pages of the entire graph. Then the computation is done iteratively. In each iteration, for each page p, find all the pages{q1, q2, …qn} that link to it. For each page qi of them, if it has m links, then the pageRank value propagated to page p is the pageRank of page qi in the previous iteration divided by m. Add all these propagated values up to be the new pageRank value of page p.

To avoid the page sinks and better model the surfing model, the damping factor c is introduced into the computation to describe the fact that the surfer could always stop following links and jump to a random page with a small probability. This also helps to guarantee the convergence of the pageRank computation.

After the pageRank computation is done on the entire collection of the pages, for a given query, first we use Vector Space Similarity algorithm to find all the pages relevant to the query, and then for each of them, compute the weighted combination of its vector similarity and pageRank value. Then the top k pages with the highest combined value will be returned as the result.

The implementation of the algorithm is straightforward. The given API provides the method to find the pages pointing to and pointed by a given page. The pageRank vector is represented as a hash table with the page names as the keys and the pageRank values as the values. First all the pageRank values are initiated to be 1/N. Then in each iteration, make a copy of pageRank vector as the source vector, and make a new pageRank vector with all the values being zero as the destination vector. Then go through all the pages in this source hash table. For each page, find all the m pages it points to, lookup the pageRank value of this page in the source vector and divide it by m, and add the result to the pageRank value of each of the linked pages in the destination vector. After this, use the damping factor to modify the destination vector. If the computation is not converged yet, the destination vector will be the source vector of next iteration.

To detect the convergence, in my implementation the difference between the source vector and the destination vector is calculated and compared to a predefined threshold. If the difference is smaller than the threshold, the computation is considered to be converged. In this implementation a very small threshold 0.00001 is used and it works pretty well in the given setting.

After the pageRank computation is done, for any given query, we just first find all the relevant pages using Vector Space Similarity algorithm implemented in project A, and get the weighted combination of the similarity values of the pages and the pageRank values to be the final values used to rank the pages.

Important observations of implementation

During the implementation one of the important observations is, compared to the Vector Space Similarity value, the pageRank value of any page is always a very small float value. The more pages in the entire collection, the smaller the pageRank values are. Because the pageRank transition matrix is a stochastic matrix, where all the pages share the total pageRank value of 1. For example, in the experiment setting, the page www.asu.edu is the one with the highest pageRank, which is only around 0.0476, and the other pages have much smaller pageRank values. At the same time, for most queries, the Vector Space Similarity values of the most similar pages are usually much higher, for example around 0.4 or even larger. So for any relevant page, its pageRank value is very small compared to its document similarity value (and we only have less than 10, 000 pages in the collection, so if it is the entire web, this gap would be much larger). The result is that if we directly take the combination of these two values, the pageRank value will have very little effect on the final result even if we have a large weight (for example, 0.9) for the pageRank value. To overcome this problem, in my implementation, after the pageRank computation is converged, the pageRank values are normalized. There could be many ways to do the normalization, and in the implementation, a simple one is used which just divide all the values by the largest value (so now www.asu.edu will have pageRank as 1). This is simple but it seems to be working well enough in the experiments.

Another implementation issue is the choice of threshold of convergence. Unlike A/H computation, the threshold in the pageRank computation should be smaller number because of the same reason stated above. The 0.00001 threshold seems to be working fine.

Comparison of pageRank and A/H computation

The computation of pageRank and A/H values are similar in the sense that they both use the link information to find the importance of the pages, and the computation are both done iteratively and converge finally.

The difference is that the A/H computation is generally done on a smaller set of relevant pages returned by Vector Space Similarity algorithm, and pageRank is done on the entire collection of pages and the values will be combined with the Vector Space Similarity values of the relevant pages of given query.

Also the pageRank model takes the random surfing into consideration which not only approaches the reality better but also makes the result more stable. Also by directly combining the Vector Space Similarity value and the pageRank value it shows user the pages which are usually both “relevant” and “important”. In the experiment, pageRank almost always performs better then the A/H method. For example for the query “Information Retrieval”, the A/H method returns the following pages:

```
Top 10 authority pages:

Rank      Authority Value          URL

--------  ----------------------   ------------------------------

0         0.35425341304919206      www.asu.edu

1         0.25267398239431327      www.asu.edu%%

2         0.22422724684322476      isa.asu.edu

3         0.2218393816185337       ame.asu.edu%%contact%%index.html

4         0.2218393816185337       ame.asu.edu%%news%%index.html

5         0.2218393816185337       ame.asu.edu%%index.html

6         0.2218393816185337       ame.asu.edu%%research%%index.html

7         0.2218393816185337       ame.asu.edu%%participate%%index.html

8         0.2218393816185337       ame.asu.edu%%

9         0.2218393816185337       ame.asu.edu%%faculty%%index.html
```

```
Top 10 Hub pages:

Rank      Hub Value                URL

--------  ----------------------   ------------------------------

0         0.31963679874160217      ame.asu.edu%%education%%index.html

1         0.2627525888597416       ame.asu.edu%%education%%courses.html

2         0.2627525888597416       ame.asu.edu%%education%%programs.html

3         0.2627525888597416       ame.asu.edu%%education%%apply.html
```

```
4          0.24068207901975816      ame.asu.edu%%contact%%index.html
5          0.2402849519166885       ame.asu.edu%%education%%faq.html
6          0.2182144420767051       ame.asu.edu%%research%%index.html
7          0.21821444207670507      ame.asu.edu%%facilities%%index.html
8          0.21821444207670507      ame.asu.edu%%news%%index.html
9          0.21821444207670507      ame.asu.edu%%participate%%index.html
```

For this particular query, most of the pages returned are not very relevant to the query, but the pageRank method returns the following pages:

```
Rank       Combined pageRank/Similarity         URL
--------   ----------------------           ------------------------------
0          0.13644245940506677      rakaposhi.eas.asu.edu%%cse494%%intro.html
1          0.0971483574461965       www.eas.asu.edu%%~cse408%%syllabus.html
2          0.08556924424706437      www.public.asu.edu%%~candan%%cv.htm
3          0.07675300464809105      ame.asu.edu%%research%%index.html
4          0.07134208147057854      aria.asu.edu%%people.htm
5          0.06467188344819207      www.eas.asu.edu%%~gcss%%wp%%index.html
6          0.06345258670464345      www.public.asu.edu%%~candan%%time%%index.html
7          0.06186486479337361      www.eas.asu.edu%%~csedept%%academic%%syllabi%%syl408.html
8          0.06123897825193348      www.eas.asu.edu%%~gcss%%introgcss.html
9          0.058674475232281635     www.fulton.asu.edu%%imes%%knowledge.html
```

Most of the pages returned are highly relevant. The same thing happens on almost all of the other test queries.

At the same time, pageRank method also out performed pure Vector Space Similarity ranking in most cases in the experiment setting. By taking the importance of pages into consideration, it avoids showing the user pages which contain almost the exactly same content with the query but actually are of no importance.

Effect of varying the weight of pageRank values

The final ranking values of the pages are the weighted combination of Vector Space Similarity value and pageRank value. The weight given to pageRank varies between 0 and 1.

In the experiment, changing the weight does affect the ranking of the results. When the weight is larger, the pages showed are more likely to be “important” pages and when the weight is smaller, the ranking is more close to the pure Vector Space Similarity ranking. For example for the given query “parking”, when the weight is 0.8, the result is:

```
Top 10 pages with combined pageRank/vector similariry value:

Rank     Combined pageRank/Similarity                  URL
--------  ----------------------          ------------------------------
0        0.10601560181582441     www.west.asu.edu%%adaff%%auxs%%parking%%map.htm
1        0.08372545716912018     www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2        0.0832823702960163      www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3        0.0624201696700884      www.east.asu.edu%%admin%%pts%%events%%index.htm
4        0.0612708345197573      asuartmuseum.asu.edu%%information%%info.htm
5        0.06073971516483965     herbergercollege.asu.edu%%museum%%information%%info.htm
6        0.0606513925966805      www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
7        0.06046343076296053     www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
8        0.0604469650572266      www.asu.edu%%dps%%pts%%maps%%visitormap.html
9        0.05996277341303742     www.asu.edu%%dps%%pts%%maps%%vmemorialunion.html
```

And when the weight is set to be 0.4, the result is:

```
Top 10 pages with combined pageRank/vector similariry value:

Rank     Combined pageRank/Similarity                  URL
--------  ----------------------          ------------------------------
0        0.3129419222832157      www.west.asu.edu%%adaff%%auxs%%parking%%map.htm
1        0.24694662410369986     www.asu.edu%%dps%%pts%%maps%%vparkingservices.html
2        0.24525467927120295     www.asu.edu%%dps%%pts%%maps%%parkingservices.html
3        0.1834445640206529      www.east.asu.edu%%admin%%pts%%events%%index.htm
4        0.17736174617319558     www.asu.edu%%dps%%pts%%maps%%studenthealthcenter.html
5        0.17709232067751185     asuartmuseum.asu.edu%%information%%info.htm
6        0.176826761000053       herbergercollege.asu.edu%%museum%%information%%info.htm
7        0.17679786067203565     www.asu.edu%%dps%%pts%%maps%%studentservicesbuilding.html
8        0.1760730466658222      www.asu.edu%%dps%%pts%%maps%%visitormap.html
9        0.17512054530954438     www.asu.edu%%dps%%pts%%maps%%studentrecreationcenter.html
```

We noticed that the top 4 or 5 pages seem to be stable. Actually in the given setting, same thing happens in most of the queries. The main reason is that for those queries, the top several pages (for example, top 4) usually have much higher Vector Similarity values than other pages in top 10, so even the weight of pageRank changes, the Vector Similarity value still dominate the final rank value thus the ranking stays stably. Of course these might not always be true because there are certain queries which are not very similar to any page, in those cases change of pageRank weight would change the ranking a lot.

Effect of varying the damping factor

In the pageRank method, the damping vector c is to model the probability of a surfer going from a page to another without following the link. In our implementation, after each iteration, the pageRank vector would be modified as:

pageRank[i] = c* pageRank[i] + (1-c)/N

In the experiment we found that when changing c in a reasonable range, it will not change the pageRank **ranking** a lot. Although the actual pageRank values might change, but we only care the ranking and it turns out that the ranking does not change that much unless c is changed very dramatically. For example in the experiment we use c = 0.4, c = 0.6 and c = 0.8, and the following is the results of query “parking decal” respectively:

```
Damping Factor 0.4
Weight of the page rank value: 0.5
Top 10 pages with combined pageRank/vector similariry value:

Rank     Combined pageRank/Similarity              URL
--------  ----------------------        ------------------------------
0        0.23532491838210753     www.asu.edu%%dps%%pts%%decals%%vendor.html
1        0.21530664214207834     www.asu.edu%%dps%%pts%%decals%%faculty.html
2        0.20913042304901652     www.asu.edu%%dps%%pts%%decals%%renewing.html
3        0.20501599005762855     www.asu.edu%%hr%%new_employee%%parking_decal.html
4        0.20165883400432996     www.asu.edu%%dps%%pts%%decals%%options.html
5        0.19982339690708753     www.asu.edu%%dps%%pts%%decals%%display.html
6        0.1900525583985174      www.east.asu.edu%%admin%%%%pts%%decal%%index.htm
7        0.17982954780886545     www.asu.edu%%dps%%pts%%decals%%howto.html
8        0.17663757525594617     www.east.asu.edu%%admin%%pts%%regulations%%index.htm
9        0.16790854021354723     www.east.asu.edu%%admin%%pts%%residences%%index.htm
```

```
Damping Factor 0.6
Weight of the page rank value: 0.5
Top 10 pages with combined pageRank/vector similariry value:

Rank     Combined pageRank/Similarity              URL
--------  ----------------------        ------------------------------
0        0.23490406028985203     www.asu.edu%%dps%%pts%%decals%%vendor.html
1        0.21488578404982284     www.asu.edu%%dps%%pts%%decals%%faculty.html
2        0.20870956495676102     www.asu.edu%%dps%%pts%%decals%%renewing.html
3        0.20459119077706495     www.asu.edu%%hr%%new_employee%%parking_decal.html
4        0.20123797591207446     www.asu.edu%%dps%%pts%%decals%%options.html
5        0.19940253881483203     www.asu.edu%%dps%%pts%%decals%%display.html
6        0.1896037062718342      www.east.asu.edu%%admin%%%%pts%%decal%%index.htm
7        0.1795157775787044      www.asu.edu%%dps%%pts%%decals%%howto.html
```

```
8         0.17618848662590025     www.east.asu.edu%%admin%%pts%%regulations%%index.htm
9         0.1674594515835013      www.east.asu.edu%%admin%%pts%%residences%%index.htm
```

```
Damping Factor 0.8
Weight of the page rank value: 0.5
Top 10 pages with combined pageRank/vector similariry value:

Rank      Combined pageRank/Similarity              URL
--------  ----------------------          ------------------------------
0         0.23532491838210753     www.asu.edu%%dps%%pts%%decals%%vendor.html
1         0.21530664214207834     www.asu.edu%%dps%%pts%%decals%%faculty.html
2         0.20913042304901652     www.asu.edu%%dps%%pts%%decals%%renewing.html
3         0.20501599005762855     www.asu.edu%%hr%%new_employee%%parking_decal.html
4         0.20165883400432996     www.asu.edu%%dps%%pts%%decals%%options.html
5         0.19982339690708753     www.asu.edu%%dps%%pts%%decals%%display.html
6         0.1900525583985174      www.east.asu.edu%%admin%%%%pts%%decal%%index.htm
7         0.17982954780886545     www.asu.edu%%dps%%pts%%decals%%howto.html
8         0.17663757525594617     www.east.asu.edu%%admin%%pts%%regulations%%index.htm
9         0.16790854021354723     www.east.asu.edu%%admin%%pts%%residences%%index.htm
```

We can see that for this particular query, the change of damping factor does not change the final ranking at all.

But a very interesting observation is that although varying damping factor (in some range) does not change the final ranking of the pages dramatically, it does influent the speed of pageRank convergence **greatly**. For example, in the above experiment, for the given convergence threshold, when c = 0.8, it takes 22 iterations to converge; when c = 0.6, it takes 11 iterations and when c = 0.4 it takes only 7 iterations to converge. The reason is that in the equation

pageRank[i] = c* pageRank[i] + (1-c)/N

the c*pageRank[i] part describe the importance of the link structure, which has very different values for different pages. And the (1-c)/N part models the "random jumping" of a surfer, which does not differentiate each page. So the result is that when c is getting smaller, the (1-c)/N part count more in the pageRank value, and the difference between pages are reduced, so the pageRank values are more evenly distributed among pages instead of propagating to some pages (but of course not absolutely stop propagating). Thus apparently it takes less iterations to converge. This also proves that the "ranking" converges faster than the "value" does, because the range of "ranking" is a discrete finite space.

Efficiency Issues

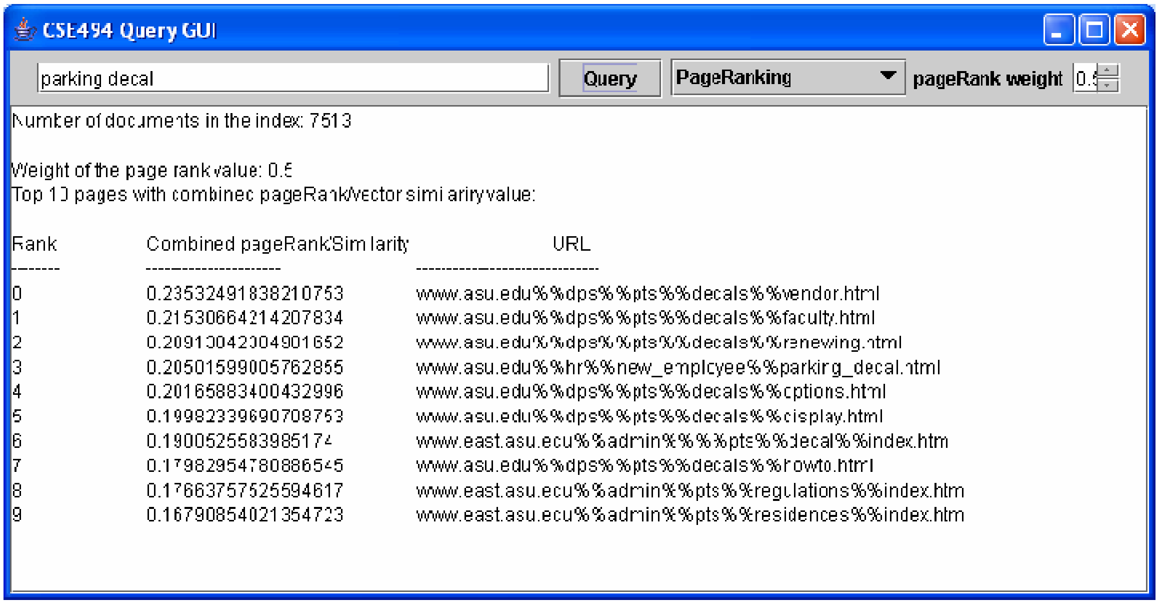
As stated earlier, pageRank computation avoids the normalization in each iteration, and the computation itself is not very complex. So what matters is the size of the collection. In my implementation, the speed is satisfactory: it only takes less than 10 seconds for the 22 iterations before convergence. Moreover, this is done only once for the entire collection and after that, the combination of the pageRank value and the Vector Space Similarity is just trivial and it takes very short time to process each query. Overall the efficiency is better than the A/H computation, as stated in report 1.

CSE494 Information Retrieval

Project B report 3 (GUI)

A GUI is implemented as required in the Extra Task. This GUI provides the interface for user to input the query, choose the method of ranking (Vector Space Similarity, A/H ranking, or combined Vector Space Similarity value and pageRank). Also it allows the user to specify the weight given to the pageRank value when using the combined similarity/pageRank value. The final result will be showed in the text area.

The following are the screenshot of the GUI during usage:



parking decal

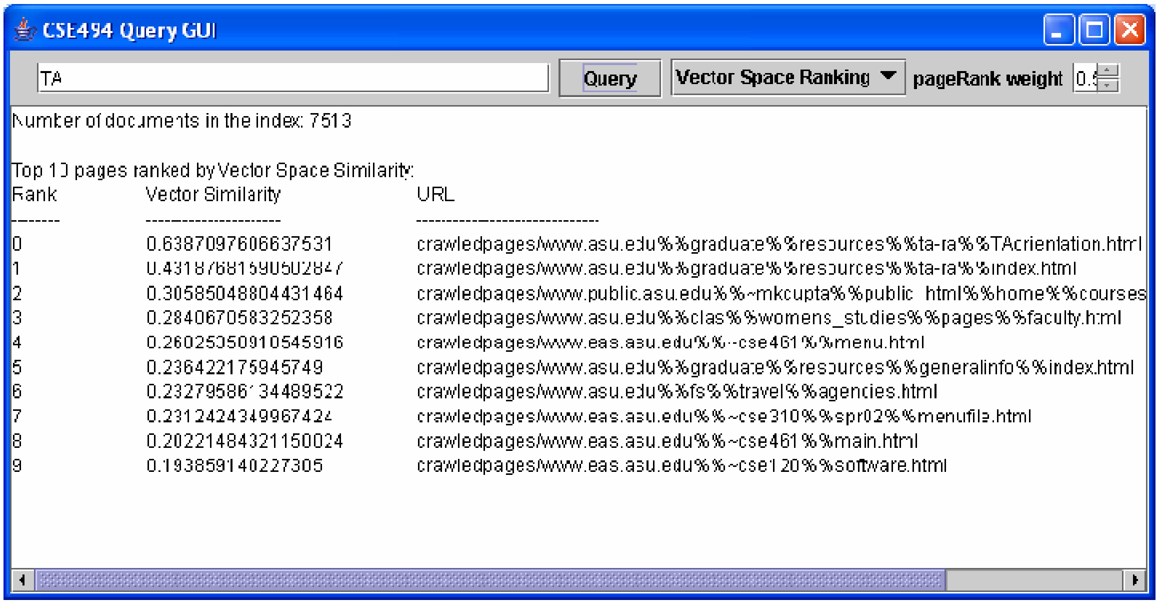
PageRanking

pageRank weight 0.5

```
Number of documents in the index: 7513

Weight of the page rank value: 0.5
Top 10 pages with combined pageRank/vector similarity value:

Rank           Combined pageRank/Similarity              URL
--------       ----------------------         ------------------------------
0              0.23532491838210753            www.asu.edu%%dps%%pts%%decals%%vendor.html
1              0.21530664214207834            www.asu.edu%%dps%%pts%%decals%%faculty.html
2              0.20913042304901652            www.asu.edu%%dps%%pts%%decals%%renewing.html
3              0.20501599005762855            www.asu.edu%%hr%%new_employee%%parking_decal.html
4              0.20165883400432996            www.asu.edu%%dps%%pts%%decals%%options.html
5              0.19982339690708753            www.asu.edu%%dps%%pts%%decals%%display.html
6              0.1900525583985174             www.east.asu.edu%%admin%%%%pts%%decal%%index.htm
7              0.17982954780886545            www.asu.edu%%dps%%pts%%decals%%howto.html
8              0.17663757525594617            www.east.asu.edu%%admin%%pts%%regulations%%index.htm
9              0.16790854021354723            www.east.asu.edu%%admin%%pts%%residences%%index.htm
```



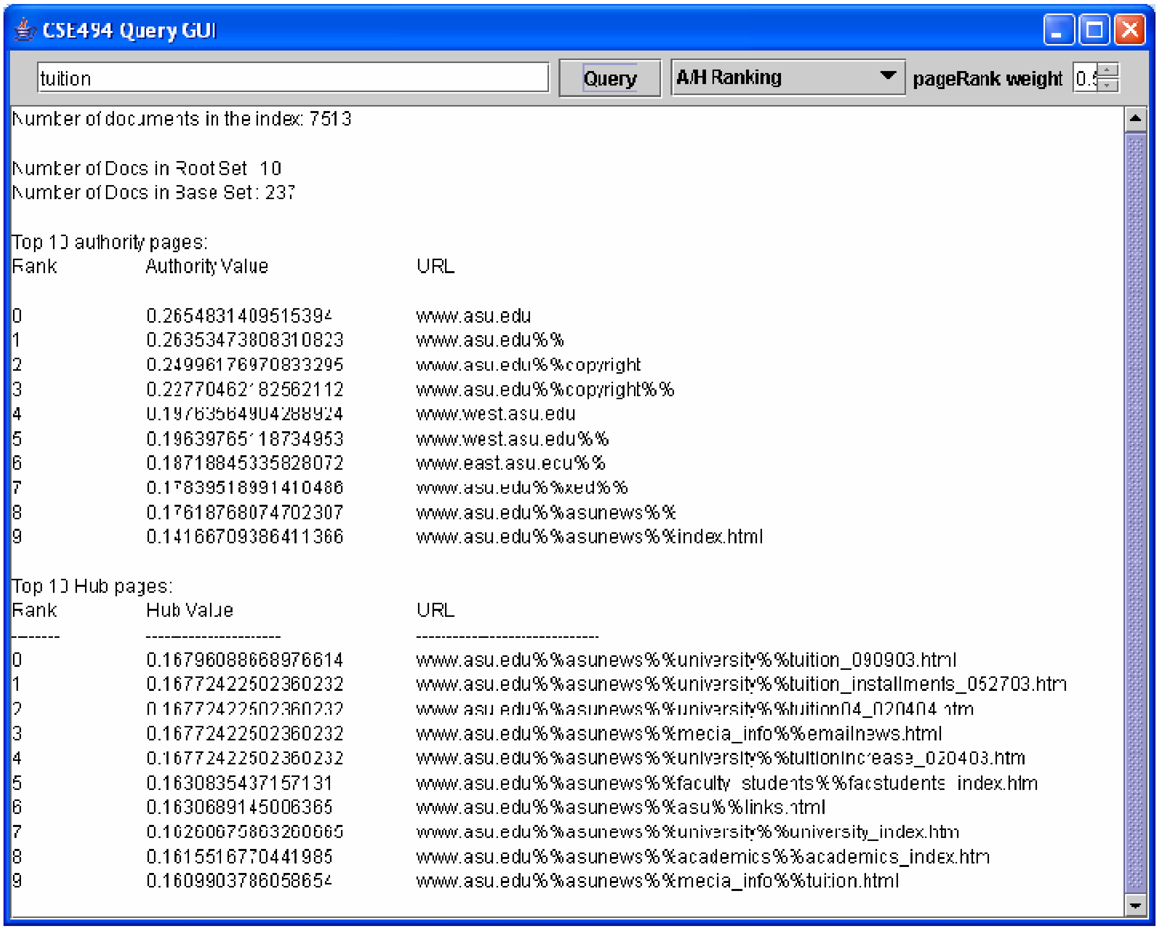
TA

Vector Space Ranking

pageRank weight 0.5

```
Number of documents in the index: 7513

Top 10 pages ranked by Vector Space Similarity:
Rank           Vector Similarity              URL
--------       ----------------------         ------------------------------
0              0.6387097606637531             crawledpages/www.asu.edu%%graduate%%resources%%ta-ra%%TAorientation.html
1              0.43187681590502847            crawledpages/www.asu.edu%%graduate%%resources%%ta-ra%%index.html
2              0.30585048804431464            crawledpages/www.public.asu.edu%%~mkgupta%%public_html%%home%%courses
3              0.2840670583252358             crawledpages/www.asu.edu%%clas%%womens_studies%%pages%%faculty.html
4              0.26025350910545916            crawledpages/www.eas.asu.edu%%~cse461%%menu.html
5              0.236422175945749              crawledpages/www.asu.edu%%graduate%%resources%%generalinfo%%index.html
6              0.23279586134489522            crawledpages/www.asu.edu%%fs%%travel%%agencies.html
7              0.2312424349967424             crawledpages/www.eas.asu.edu%%~cse310%%spr02%%menufile.html
8              0.20221484321150024            crawledpages/www.eas.asu.edu%%~cse461%%main.html
9              0.193859140227305              crawledpages/www.eas.asu.edu%%~cse120%%software.html
```



tuition

0.5

```
Number of documents in the index: 7513

Number of Docs in RootSet: 10
Number of Docs in Base Set: 237

Top 10 authority pages:
Rank            Authority Value              URL

0               0.2654831409515394           www.asu.edu
1               0.26353473808310823          www.asu.edu%%
2               0.24996176970833295          www.asu.edu%%copyright
3               0.22770462182562112          www.asu.edu%%copyright%%
4               0.19763564904288924          www.west.asu.edu
5               0.19639765118734953          www.west.asu.edu%%
6               0.18718845335828072          www.east.asu.ecu%%
7               0.17839518991410486          www.asu.edu%%xed%%
8               0.17618768074702307          www.asu.edu%%asunews%%
9               0.14166709386411366          www.asu.edu%%asunews%%index.html

Top 10 Hub pages:
Rank            Hub Value                    URL
--------        ---------------------        ------------------------------
0               0.16796088668976614          www.asu.edu%%asunews%%university%%tuition_090903.html
1               0.16772422502360232          www.asu.edu%%asunews%%university%%tuition_installments_052703.htm
2               0.16772422502360232          www.asu.edu%%asunews%%university%%tuition04_020404.htm
3               0.16772422502360232          www.asu.edu%%asunews%%mecia_info%%emailnews.html
4               0.16772422502360232          www.asu.edu%%asunews%%university%%tuitionincrease_020403.htm
5               0.1630835437157131           www.asu.edu%%asunews%%faculty_students%%facstudents_index.htm
6               0.1630689145006365           www.asu.edu%%asunews%%asu%%links.html
7               0.16260675863260665          www.asu.edu%%asunews%%university%%university_index.htm
8               0.1615516770441985           www.asu.edu%%asunews%%academics%%academics_index.htm
9               0.1609903786058654           www.asu.edu%%asunews%%mecia_info%%tuition.html
```