

Spring 2013

Intelligent Personalized Searching

Wing Lau

San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Recommended Citation

Lau, Wing, "Intelligent Personalized Searching" (2013). *Master's Projects*. 292.

DOI: <https://doi.org/10.31979/etd.432s-cck3>

https://scholarworks.sjsu.edu/etd_projects/292

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Intelligent Personalized Searching

A Project Report (CS298)

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Wing Lau

April 2013

© 2013

Wing Lau

ALL RIGHTS RESERVED

The Designated Project Committee Approves the Writing Project Titled

Intelligent Personalized Searching

By

Wing Lau

APPROVED FOR THE DEPARTMENT OF COUMPUTER SCENCE

SAN JOSÉ STATE UNIVERSITY

April 2013

DR. Tsau Young Lin

Date

Department of Computer Science

Dr. Soon Tee Teoh

Date

Department of Computer Science

Tony Kung

Date

Director Program Management

TIBCO Software Inc.

Abstract

Search engine is a very useful tool for almost everyone nowadays. People use search engine for the purpose of searching about their personal finance, restaurants, electronic products, and travel information, to name a few. As helpful as search engines are in terms of providing information, they can also manipulate people behaviors because most people trust online information without a doubt. Furthermore, ordinary users usually only pay attention the highest-ranking pages from the search results. Knowing this predictable user behavior, search engine providers such as Google and Yahoo take advantage and use it as a tool for them to generate profit. Search engine providers are enterprise companies with the goal to generate profit, and an easy way for them to do so is by ranking up particular web pages to promote the product or services of their own or their paid customers. The results from search engine could be misleading. The goal of this project is to filter the bias from search results and provide best matches on behalf of users' interest.

Acknowledgement

I would like to express my greatest gratitude to professor Lin for his endless help and valuable advice for this project. I have worked on this project for about a year, and during this time, Professor Lin gave me creative advice to solve any encountered problems. I would also like to thank Professor Silver, whom contributed a lot to this project. Professor Silver shared valuable insights as a user for this project. He also provided test cases and helped with the verification of the results. Finally, I would like to thank my committee members, Professor Teoh and Mr. Kung, for their time and support.

Contents

Abstract.....	4
Acknowledgement.....	5
Contents	6
List of Tables.....	7
List of Figures	8
1 Introduction	9
2 Theory behind this project	10
2.1 Introduction	10
2.2 LSI	10
2.3 Term Frequency-Inverse Document Frequency (TFIDF)	16
3 Implementation	17
3.1 Instruction	17
3.2 Previous approaches.....	18
3.3 Implementation details.....	19
4 Test Cases and Result Analysis.....	27
4.1 Test Result.....	27
4.2 Result Analysis	41
5 Conclusion	42
6 Future Work	42
References.....	43

List of Tables

Table 1: Original Term * Document matrix	11
Table 2: Matrix with LSI processed data.....	14
Table 3: Matrix with original data	14
Table 4: Document vector value after LSI process	15
Table 5: Document similarity value against query.....	16
Table 6: Test cases	27
Table 7: Result of test case 1	27
Table 8: Keywords of test case 1.....	28
Table 9: Result of test case 2	29
Table 10: Keywords of test case 2	30
Table 11: Result of test case 3.....	31
Table 12: Keywords of test case 3	32
Table 13: Result of test case 4.....	32
Table 14: Keywords of test case 4	33
Table 15: Result of test case 5	34
Table 16: Keywords of test case 5	35
Table 17: Result of test case 6.....	36
Table 18: Keywords of test case 6	37
Table 19: Result of test case 7.....	37
Table 20: Keywords of test case 7	38
Table 21: Result of test case 8.....	39
Table 22: Keywords of test case 8	40

List of Figures

Figure 1: Project outline.....	17
Figure 2: Source code query against Yahoo	20
Figure 3: Yahoo return page	21
Figure 4: Price table for Bing search.....	22
Figure 5: Price table form Google Search	23
Figure 6: Main function for LSI	24
Figure 7: The function finds the keywords based on TFIDF value	25
Figure 8: Web interface of this application.....	26

1 Introduction

This project was initiated by Dr. Tsau Young Lin and Dr. Steven Silver, and the goal is to provide biased-free search results to the users. Drs. Lin and Silver have noticed the problem of biased search results because the top ranking search results are always the commercial web sites.

The deeper problem is that search engine providers have expanded their domain to more than just a search engine provider. For example, Google has been growing in advertising platform, software development, hardware development, game and app development, mobile operating system development, web browser development, video hosting, blog platform, social network and most recently an e-commerce site. With the goal of generating more profits, search engine providers manipulate the search results by ranking up its own services or products. In addition to Google, Yahoo and Bing are using the same strategy to promote their services and products.

To get users the bias-free search results without building a huge knowledge database, we reuse the search results return from search engine. To accomplish this, three different steps are needed: First, we obtain the search results from search engine. Second, we reshuffle the return in new ranking, and then third, we present the results to the users in a new ranking. In theory, this method works for all the search engine providers, but there is the resource issue. This project will use Yahoo as the search engine provider even though Google is preferred. (In Later section, I will explain why we choose to use Yahoo instead of Google as our search engine.)

We have used several approaches for this project and the selected approach is using Term Frequency-Inverse Document Frequency (TFIDF) and Latent Semantic Indexing (LSI) together. This approach provides good results. The test cases for this project are the eight queries provided by Dr. Steven Silver. They are involved in technical term, product search, and life style query.

In later section of this document, the detail information will be provided pertaining to project implementation, test cases, and the theories behind the project. For reader to understand the project better, tutorials about TFIDF and LSI will be given in the later section as well.

2 Theory behind this project

2.1 Introduction

The two most import algorithms behind this project are Latent Semantic Indexing (LSI) and Term Frequency-Inverse Document Frequency (TFIDF). LSI is also known as Latent Semantic Analysis (LSA). As mentioned earlier, few approaches have been tried but failed to generate reasonable output. One of the reasons is that stop-words and hidden information were not handled well. Hidden information means higher order co-occurrence. And LSI and TFIDF are the solutions for these problems.

2.2 LSI

Latent Semantic Indexing is a method that projects users' queries and documents into semantic dimensions spaces, a row colon matrix. Each document and users' query is represented by a vector, which is usually a column vector. From computing the cosine as the similarity value of each pair of vectors, one can find out how close they are related.

Latent Semantic Space is different from original document space. Latent Semantic Indexing is the application of Singular Value Decomposition. By applying the SVD method to the original document space, the original document term matrix can be represented in lower dimension vector spaces. Furthermore, it can find the "latent" semantic relationship and reduce unnecessary noise.

In order to fully understand LSI, we must understand SVD. Professor Gene Golub developed SVD in 1965. One of his goals was to determine the singular values and pseudo-inverse of a matrix, to compute the rank of matrix by counting the number of

nonzero singular values, and to expose hidden properties and the features of the matrix under SVD.

SVD is based on the following equation $A = USV^T$. The columns of U are the eigenvectors of the AA^T matrix and the columns of V are the eigenvectors of the $A^T A$ matrix. S is the matrix composed from the singular values of A. Matrix A can be reconstructed with dimensionality reduction by restricting S with less singular values.

The following is an example from the tutorial posted on www.miislita.com by Dr. E. Garcia. The example has three simple documents and clearly illustrates the power of SVD. By applying SVD, the hidden information is shown in the dimension reduction matrix.

The three simple documents and a query:

- d1: Shipment of gold damaged in a fire.
- d2: Delivery of silver arrived in a silver truck.
- d3: Shipment of gold arrived in a truck.
- q: gold silver truck

	d1	d2	d3
a	1	1	1
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
in	1	1	1
of	1	1	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

Table 1: Original Term * Document matrix

Referring to equation $A = USV^T$

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad U = \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2625 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & -0.4078 \end{bmatrix}$$

$$S = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix} \quad V = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

The purpose of LSI is not to reconstruct the original matrix from the decomposition; instead, the goal is to keep the largest k singular values. From that, users can find more information, especially hidden information. Even better than that, the reconstructed matrix provides information with less noise compared to original matrix.

What is the best value for k ? There is no magic bullet for this question. It must be determined experimentally since there is a lot of variance for each different set of documents. Some studies suggest that for a huge amount of documents, k is better set between 100 and 200. Others suggest that k is to be set between 100 and 500. It is inconclusive what the best is without considering the data.

For the tutorial, k is set as 2, which is called Rank 2 Approximation. Basically, it keeps the first two columns of U and V , and the first two rows and columns of S .

$$A \approx A_2 = U_2 S_2 V_2^T$$

$$V_2 = \begin{bmatrix} -0.4945 & 0.6492 \\ -0.6458 & -0.7194 \\ -0.5817 & 0.2469 \end{bmatrix}$$

$$U_2 = \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & 0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3749 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \quad S_2 = \begin{bmatrix} 4.0989 & 0.0000 \\ 0.0000 & 2.3616 \end{bmatrix} \quad V_2^T = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \end{bmatrix}$$

	d1	d2	d3
a	0.9662	0.9850	1.0453
arrived	0.3003	1.1328	0.5974
damaged	0.6659	-0.1478	0.4478
delivery	-0.1476	0.9347	0.1982
fire	0.6659	-0.1478	0.4478
gold	1.1140	0.0506	0.8473
in	0.9662	0.9850	1.0453
of	0.9662	0.9850	1.0453
shipment	1.1140	0.0506	0.8473
silver	-0.2955	1.8692	0.3960
truck	0.3003	1.1328	0.5974
Totals	6.6159	7.8301	7.5151

Table 2: Matrix with LSI processed data

Table 3: Matrix with original data

	d1	d2	d3
a	1	1	1
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
in	1	1	1
of	1	1	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1
Totals	7	8	7

$$6.6159 + 7.8301 + 7.5151 = 21.9611$$

$$7 + 8 + 7 = 22$$

$$\text{Net Noise} = 22 - 21.9611 = 0.0389$$

Comparing the two tables above, they have very similar values. The differences are attributed to LSI, based on the co-occurrences of those terms in different documents. It illustrates the power of LSI, which allows for the recovery of the hidden information. In this tutorial, the words “arrived” and “truck” did not appear in d1. However, they co-

occur with the stop-words “a”, “of”, and “in” in all three documents. This is why their weight is 0.3003 in d1.

The other feature of LSI is to reduce the noise. Some of the words are not significant for document lookup or classification, and their weight may affect the correctness of the results. LSI helps compensate by increasing the weight of some words while lowering the weight for others. By doing so, the total weight may be smaller than the total weight of the ordinal matrix. LSI does so by dimension reduction. In the above example, LSI only keeps the two biggest singular values, ignoring the third one. As the result, there are a total of 0.0389 net noise reductions. It may not seem significant, but in real cases, it makes a huge difference as there will a huge amount of documents.

There is another way to demonstrate LSI. As M.W. Berry, et al. described in their paper, the document can be described in a vector space as this formula $d = d^T US^{-1}$ where d is the original vector of the document. At the same time, the user query can also be treated like a document. To find its vector, the same formula is used. ($q = q^T US^{-1}$)

From the tutorial, the query is “gold silver truck”. By applying the value to the formula, the following values are produced.

Table 4: Document vector value after LSI process

	Vector Value
d1	(-0.4945, 0.6492)
d2	(-0.6458, -0.7194)
d3	(-0.5817, 0.249)
q	(-0.2140, -0.1821)

One of the ways to find the similarities between the query and each document is to compute the cosine values between the query vector and the document vectors. And that is the method used in this project.

Table 5: Document similarity value against query

	Similarity value (cosine)	Ranking
d1	-0.0541	3
d2	0.9910	1
d3	0.4478	2

By now, it should be clear how this project is going to reshuffle the results in the new ranking. As the above tutorial demonstrated, LSI is good at finding the hidden relationship and reducing the noise.

2.3 Term Frequency-Inverse Document Frequency (TFIDF)

The section above gave a detail tutorial about how powerful LSI is for finding the similarity, making it too perfect for the need to make any changes. However, LSI does not handle the stop-words very well, which is why TFIDF is introduced in this project. TFIDF is the product of two statistics, term frequency and inverse document frequency.

$$tfidf(t, d) = tf(t, d) \cdot idf(t, D)$$

The first term of the equation is easy to understand. Its value increases as the number of times the term appears in a document.

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

The equation above is for Inverse Document Frequency. The numerator is the number of total documents, and the denominator is the number of documents that contains term t . If the term appears in most of the documents, the value of idf becomes very small.

TFIDF is a numerical statistic, which reflects how to import a word in a document within a collection of documents or corpus. The idea behind this is that if some terms appear too often in the document, these terms become insignificant since they are too

abundant to stand out. The point is that common words are not the keywords. For example, “you”, “are”, “a”, “and”, nor “the” are the keywords as they are quite common in most documents. For text mining or classification, the weight of these common words should be considered to be small.

3 Implementation

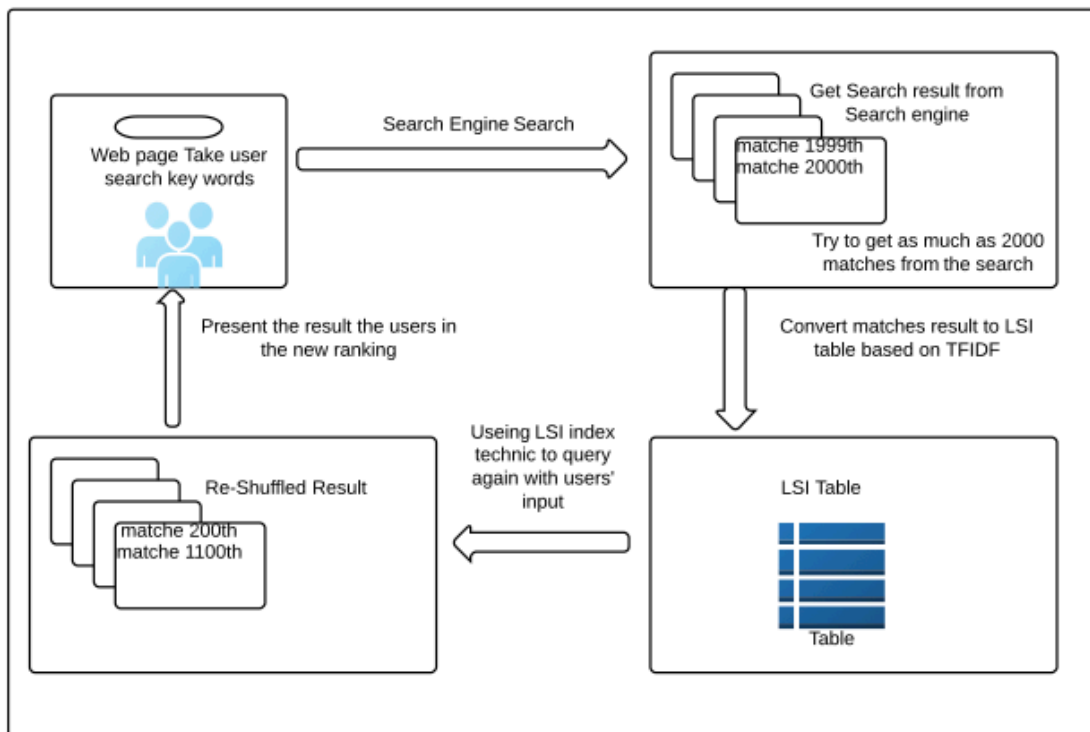


Figure 1: Project outline

3.1 Instruction

The outline of the project is shown in figure 1. The main idea is to provide a user interface to allow users enter their queries, which then generate as many matches as it can from the search engine. The results are then reshuffled and presented in the ranking based on users' interest.

This project is divided into two parts. The first part is to get the results from the search engine. The second part is to use LSI to build a table for the purpose of re-shuffling the results and presenting the results to the user in the new non-biased ranking.

3.2 Previous approaches

Before further explaining the detail of the final implementation of this project, it is important to know about a few approaches that had been tried but failed to produce good results.

The first approach is to find the word count of the keywords from users' query in the return matches (The whole page of each match return by search result). As described earlier, this is not the proper solution since the "latent" information is missing. Remember that without applying LSI, the hidden information cannot be located.

The second approach is to randomly select the results for the search engine returns. Search engine usually lists ten results per page, and there are at least hundreds of pages for a single query. If the goal is to show the users the result without bias, then randomly selecting results from the pages seems like a quick and good solution. However, the result is neither promising nor stable because results are random in nature and some returns from search engine are not the good ones. This solution may be quick but it is not a good solution. We have seen some totally unrelated contents appear in the top 10 list from this approach.

The third approach is very similar to the final approach. The difference is that the content used to build the LSI table is from the abstract of the search results, not the whole document of each match. The returns from search engine usually have three essential parts: a short tile, a link that points to the whole content, and an abstract that describes about what the linked document is about. For performance reasons, we reshuffled the return based on the abstracts only and the result is not that good. The reason may be that the abstract is too short to provide enough words to build up a good LSI table. Also, the

abstract may not contain enough keywords. From the experience of running eight different queries, we conclude this approach is not a good way to reshuffle the results.

We determined the result generated by those three approaches is not good because they have a common problem. Some of matches from those three approaches have very high ranking but they are not close related to users' queries.

3.3 Implementation details

The final approach is to let the application obtain queries from users and do the search against a search engine. By parsing each page of the returns, the application can get the short title, html link, and abstract for each result. After getting the link, the application fetches the whole content pointed by the html link. The application tries to fetch as many as 2000 documents, at which time it can either stop or continue until there are no more matches that the search engine can return. After all the documents are obtained, the application builds a LSI table based on the documents, and from that, it calculates the similarity from which the application reshuffles and returns from the search engine. The final step is to present the results to the users.

This project is implemented in Python, which is a well known program language used especially for mathematic calculation and statistical analysis. Python has many useful libraries and packages and the two main packages used for this project are Simserver and Gensim. These packages will be discussed in more details in a later section about the code.

The first part of the application is to get the search results from a search engine. The application does so by making the Representational State Transfer (REST) calls to the search engines. Basically, the query input is embedded in the uniform resource locator (URL). When the call is made by sending a HTTP request, the search engine will parse the parameters along with the URL and return the contents based on the parameters.

```

12 def yahoosearch(q, num):
13     Ldesc=[]
14     Lurl=[]
15     Ltitle=[]
16     url2 = "http://search.yahoo.com/search?b=%s&p=%s"
17     ranIndex = num * 10 + 1
18     url2a = url2 % (ranIndex,q)
19     print url2a
20     r= requests.get(url2a)
21
22     soup = BeautifulSoup(r.text)
23     result = soup.contents[1].findAll('div',{'class':"res"})
24
25     numRequL = len(result)
26
27     if numRequL == 0:
28         print "--- "+ str(numRequL)
29     for t in result:
30         try:
31             titleTab = t.div.h3.a
32             titleStr = str(titleTab)
33             url = str(titleTab['href'])
34             title = nltk.clean_html(titleStr)
35             desc = t.find(attrs={'class':'abstr'})
36             Ldesc.append(str(desc.get_text()))
37             Lurl.append(url)
38             Ltitle.append(title)
39
40         except:
41             pass
42
43     returnL=[]
44     returnL.append(Ldesc)
45     returnL.append(Lurl)
46     returnL.append(Ltitle)
47     returnL.append(numRequL)
48
49
50
51
52     return returnL
53

```

Figure 2: Source code query against Yahoo

The image above is the basic function to query against Yahoo. The code in line 16 is to generate the URL for the querying. The REST call application programming interface (API) from Yahoo search takes two parameters: the starting page and the search keywords. For example, URL “http://search.yahoo.com/search?b=1&p=iphone” and “http://search.yahoo.com/search?b=11&p=iphone” are the URL call for searching

“iphone” from Yahoo search. The first URL will return the first ten matches for the first page whereas the second URL returns the eleventh to twentieth matches for the second pages. The code in line 20 makes call to get the whole content of the whole page returned by Yahoo.

One notable mention is that our application tried to call this function 200 times, and it stopped calling either when it made 2000 calls or when it detected no more matches that the search engine could find. The detection is handling in line 25 to line 27. From the test cases, it appears that Yahoo usually only returns about 500 matches.

The good thing about the yahoo return is that the matches in the return are presented in a well-defined html tabs and cascading style sheet (CSS) classes. For looking at the content inside of tabs with specific class, the application can find all the necessary information such as the short title of content, URL link to the whole content, and the abstract of the content. In the code from line 23 to line 42, it is for getting necessary information from parsing of the return page. In the return from Yahoo, each match is embedded in a div tab with class being set to “res”. By parsing those tabs in detail, the application can get all the related values.

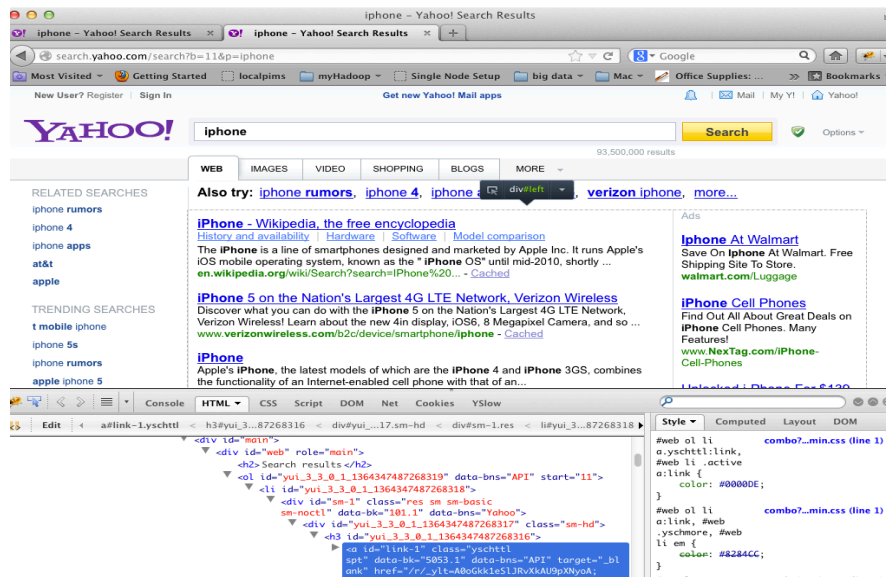


Figure 3: Yahoo return page

At this point, one may be curious why the test scenario used Yahoo instead of Google or Bing as the search engine. The main reason is funding. At the implementation time, unlimited REST calls worked only when against Yahoo, not Bing nor Google. After making about two or three REST calls to query from Google, Google detected the activities and blocked the call from my machine for about 20 minutes. Google and Bing provide and preserve search engine web service to their paid customers, but there is not enough funding allotted for this project to use their services.

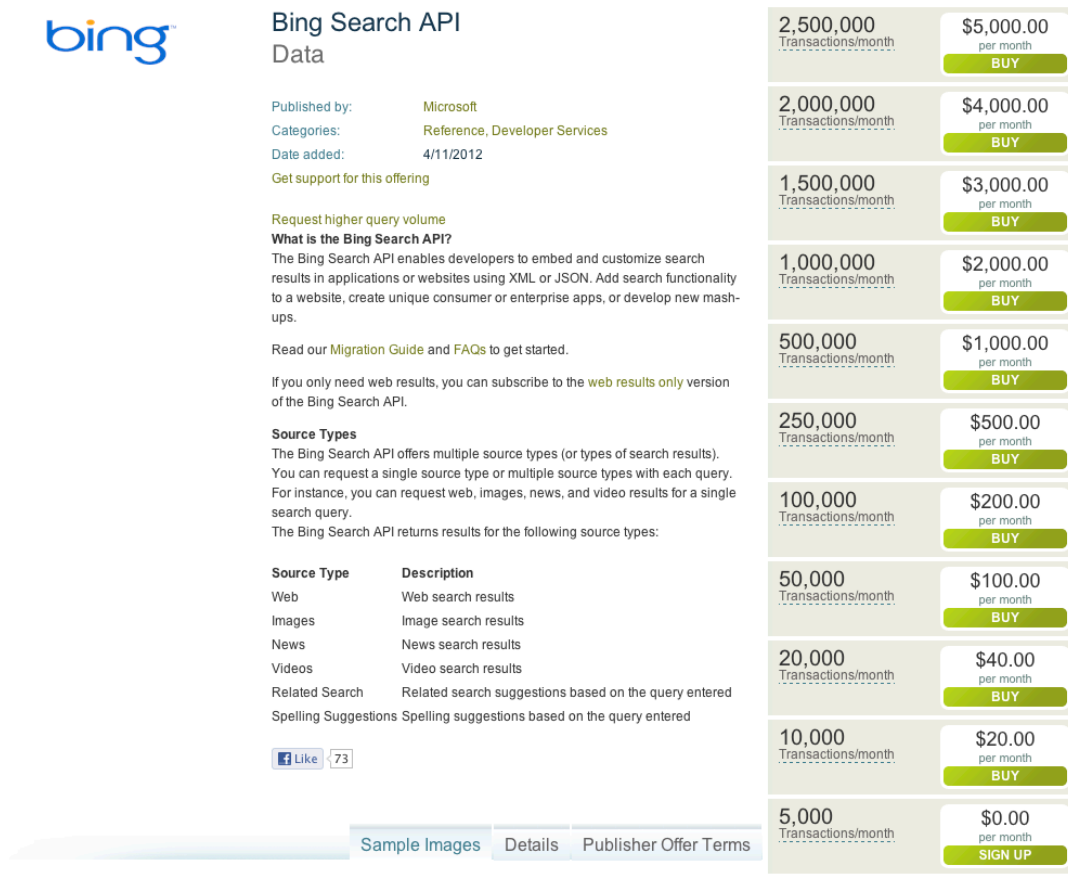



Figure 4: Price table for Bing search



Custom Search Help

Help home

Google Site Search (paid)

Google Site Search

Sign up for Google Site search

Google Site Search pricing

Non-profits

Query: Definition

About Google Site Search

Google Site Search vs Google Custom Search

Custom Search vs Google.com

Google Site Search vs Google.com

Business group

Renew Google Site Search license

Help articles › [Google Site Search \(paid\)](#)

Google Site Search pricing

Google Site Search has the following options available, depending on your estimated query traffic on your website:

Google Site Search pricing

Search query limit (annual)	Pricing (annual)
20,000	\$100
50,000	\$250
150,000	\$750
500,000	\$2,000
500,000+	Contact sales

You can sign up at <http://www.google.com/sitesearch> and pay online using Google Checkout.

Figure 5: Price table form Google Search

The second part of this application is to apply LSI on the documents obtained from the first part and calculate the similarity values between user's query and each document. The main function of this application is based on library Simserver and Gensim. Notably, Simserver is also implemented based on Gensim. In my code, Simserver is used for LSI function and Gensim is used for finding the most significant keywords from the top ranking documents. I use those keywords as the help to prove the result.


```

228 def queryFromPickleDoc(qNum, query):
229     pickleFile = "readDoc"+ str(qNum)+ '.pkl'
230     pf = open(pickleFile, 'rb')
231     texts = pickle.load(pf)
232     pf.close();
233
234     #create corpus
235     corpus = [{'id': 'doc_%i' % num, 'tokens': utils.simple_preprocess(text)} for num, text in enumerate(texts)]
236
237     print strftime("%Y-%m-%d %H:%M:%S", gmtime()) + " done generate corpus"
238     serverInstance = '/tmp/serverdoc'+str(qNum)+'/'
239
240     #create service
241     service = SessionServer(serverInstance)
242     print "done set up index"
243
244     #create training set
245     service.train(corpus, method='lsi')
246     print "done covert to lsi"
247     print strftime("%Y-%m-%d %H:%M:%S", gmtime()) + "done set up model"
248
249     #create index
250     service.index(corpus)
251     print strftime("%Y-%m-%d %H:%M:%S", gmtime()) + " end index search"
252     doc = {'tokens': utils.simple_preprocess(query)}
253     sims = service.find_similar(doc, max_results=50)
254     print strftime("%Y-%m-%d %H:%M:%S", gmtime()) + " start end"
255     pickleFile = "simDoc"+ str(qNum)+ '.pkl'
256     output = open(pickleFile, 'wb')
257     pickle.dump(sims, output)
258     output.close()
259

```

Figure 6: Main function for LSI

The image above is the main function used to apply LSI to the search results. There are two parts to the function displayed above. First, line 229 to line 250 is to set up the LSI tables. Line 252 and onward is to find the vector value for the user query, from which the function can find the cosine value against each document. By the end of this function, we can find the similarity between users' queries and each document. In line 253, the code defines only for those similar matches that are the first few in the return and the rest will be ignored. Since this application is just a prototype, it does not need to show all the matches.

```

374 def convertDocsToTfidf(qNum, selectedDocs, numOfWork):
375
376     wordCount = defaultdict(int)
377     wordHash = defaultdict(int)
378     pickleFile = "readDoc"+ str(qNum)+ '.pkl'
379     pf = open(pickleFile, 'rb')
380     texts = pickle.load(pf)
381     pf.close();
382
383     texts = [utils.simple_preprocess(doc) for doc in texts]
384
385     dictionary = corpora.Dictionary(texts)
386     corpus = [dictionary.doc2bow(text) for text in texts]
387
388     tfidfModle = models.TfidfModel(corpus)
389     tfidfCorpus = tfidfModle[corpus]
390     tfidfCorpus = list(tfidfCorpus)
391
392     for item in selectedDocs:
393         item = int(item)
394         doc = tfidfCorpus[item]
395         sortedItem = sorted(doc, key=lambda tup:tup[1], reverse =True)
396         i = 0
397
398         #print sortedItem
399         for item in sortedItem:
400
401             if i >= numOfWork:
402                 break
403             else:
404                 if item[0] in wordCount:
405                     wordCount[item[0]] += item[1]
406                 else:
407                     wordCount[item[0]] = item[1]
408
409             i = i + 1
410
411     sortedKey = sorted(wordCount, key = wordCount.get, reverse = True)
412     print "key words"
413     print
414     j = 0;
415     for key in sortedKey:
416         if j < 20:
417             print str(dictionary[key]) + ' ---> ' + str(wordCount[key])
418             wordHash[dictionary[key]] = wordCount[key]
419         else:
420             break
421         j = j + 1
422

```

Figure 7: The function finds the keywords based on TFIDF value

In reference to the tutorial, LSI reconstructs term by documenting matrix into a new dimensional reduced matrix by SVD. The tutorial shows that each document is represented by a vector in the matrix. By looking at an individual vector in the matrix, we can find what the keywords are for each document, which are the words that have high

weight values. To verify the theory and the implementation of the project, we used the function displayed above to get the key words from the top ranking documents. From this, we can determine if the keys words are closely related to the users' query.

The title for this project is Intelligent Personalized Searching. To make it more personalized, a new feature is introduced to find similar contents. In some situations, users find a very interesting match from their search results and would like to view other closely related matches. As the application already has the LSI table, to find the similarity among documents will not consume more computing resource. Furthermore, since the input is changed from a query vector to a document vector, we will expect different result; document has much more words than a query has. As the result, finding documents that are closely related to those that users like may provide users more helpful information. In the web page interface, there is a button for users to find similar matches to those of the attached match.

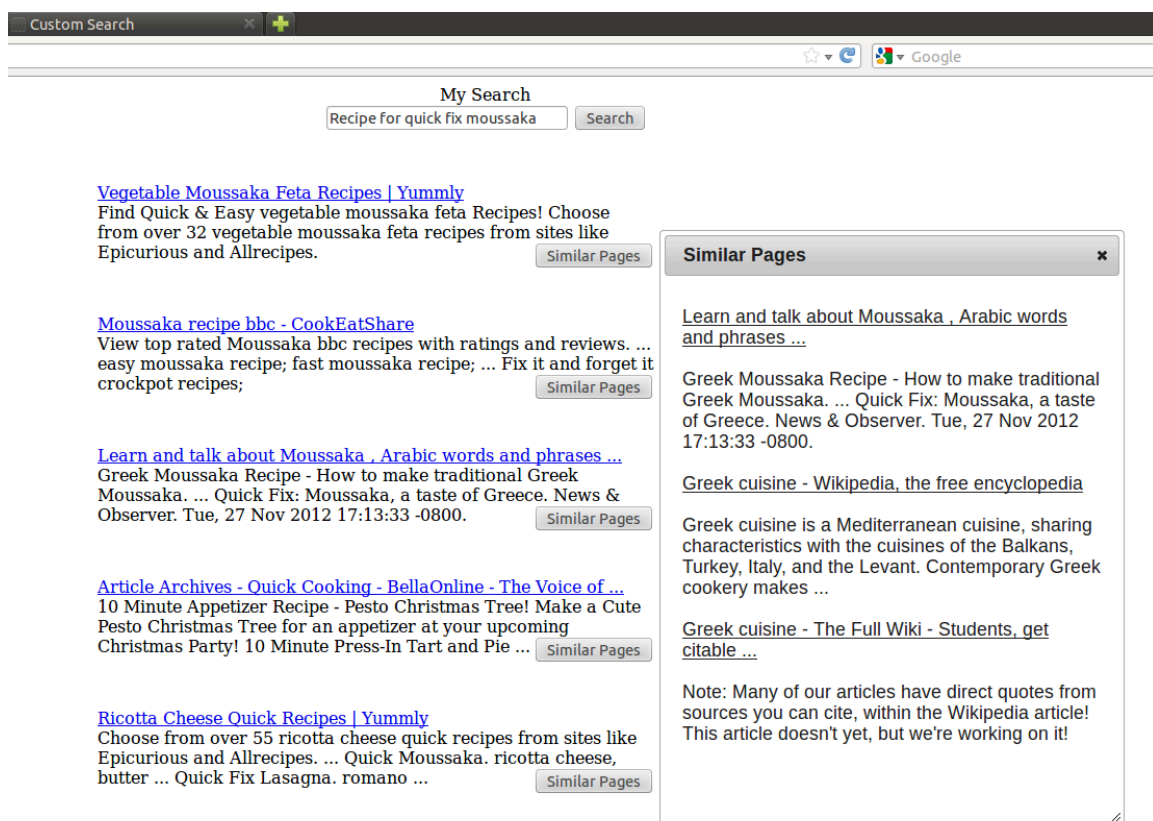


Figure 8: Web interface of this application

4 Test Cases and Result Analysis

To test the application, we have run eight queries against our application. Dr. Steven Silver provided queries that involved technical terms, product search, and lifestyle search. The result was promising, as shown by the section below. The application provides the top 50 LSI ranking matches for each query, but only the top ten matches will be listed. The weight value of each keyword is calculated from the sum of its weight in the top 50 ranking documents. Both the top keywords and the weight value are shown in the table below.

4.1 Test Result

Table 6: Test cases

	Query
1	Agent based models in dynamic and social networks
2	Tablet with HDMI and external keyboard
3	Shortest travel time Los Angeles to Taipei
4	Highest ranking mandarin cuisine within 20 miles of Santa Clara, California
5	Current capabilities of artificial intelligence for natural language
6	Software for non-linear function fitting
7	Recipe for quick fix moussaka
8	Online quality gemstones

Table 7: Result of test case 1

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	2	0.632	Agent - based model - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/Agent-based_model An agent-based model ... which was discussed in his paper Dynamic Models of Segregation in 1971. ... especially to models of social networks, ...
2	32	0.529	Journal of Simulation - Tutorial on agent - based modelling and ...	http://www.palgrave-journals.com/jos/journal/v4/n3/full/jos20103a.html For dynamic networks, ... Sakoda (1971) formulated one of the first social agent-based models, the Checkerboard Model, which relied on a cellular automaton.
3	12	0.512	Agent Based Models with Anylogic - Coensys - PLM, PDM ...	http://www.coensys.com/agent_based_models.htm In case of large number of agents with dynamic connections (such as social networks) ... To add synchronization to your agent based model. 1.

4	31	0.441	Formation of Economic and Social Networks (Tsfatsion)	http://www2.econ.iastate.edu/tesfatsi/netgroup.htm Agent-Based Computational Economics ... "Dynamic Social Networks Promote Cooperation in Experiments with ... dynamic social networks; multi-agent network models; ...
5	7	0.413	Networks : Agent - Based Computational Economics (Tsfatsion)	http://www.econ.iastate.edu/tesfatsi/anetwork.htm ... Agent-based models typically involve large numbers of interacting individuals with ... dynamic social networks; multi-agent network models; group ...
6	14	0.412	Agent - Based Modeling and Simulation Researchers: R Agent ...	http://www.agent-based-models.com/blog/researchers/r/ ... use of agent-based modeling to study the dynamics of ... reasoning and social behavior; neural network models of ... of Agent-Based Models in Social ...
7	25	0.39	Business and Management Modeling : Agent - Based Computational ...	http://www.econ.iastate.edu/tesfatsi/abusiness.htm ... Supply Networks: An Agent-Based ... in urban dynamics. Different types of dynamic models are ... Social simulation and behavioural dynamics; ...
8	57	0.385	Individual- Based Models	http://www.red3d.com/cwr/ibm.html Individual-based models are also known as entity or agent based models, ... network might be based on individual models ... dynamic recreation behavior, social ...
9	9	0.375	Agent - Based Modeling : Social Sciences Agent - Based Models	http://www.agent-based-models.com/blog/resources/social-sciences/ A list of websites that use agent-based models in the social ... research line is on modeling the interdependent dynamics of social norms and social networks.
10	129	0.305	NETSCI 09 list of submissions	http://pilastro.phys.uniroma1.it/netsci/docs/submission_show_all.cgi.html Abstract. The sustained increase in different forms of electronic interaction over the last decade has led to the emergence of a number of electronic and visible ...

Table 8: Keywords of test case 1

Keyword	Weight
agent	4.408676615
the	3.140417871
simulation	2.979673397
of	2.73409476
networks	2.611439415
agents	2.381951752
modeling	1.867237768
and	1.795130134
model	1.76167145
abm	1.64266745
models	1.566597833
computational	1.211249784
ecd	1.129145774

travel	1.121418733
jq	1.064596869
network	1.044270728
google ad	0.991229784
abstract	0.980327514
systems	0.963790296
netlogo	0.882659738

Table 9: Result of test case 2

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	4	0.401	Android Tablet External Keyboard ,Buy Quality Android Tablet ...	http://www.alibaba.com/showroom/android-tablet-external-keyboard.html Android Tablet External Keyboard, ... Android 4.04 512MB+4G 4.Wifi wireless,3400mAh battery 5.HDMI jack... Type: Tablet PC Tags: ...
2	172	0.351	Android Tablet » Search Results » Android Tablet Hdmi	http://thepctablet.info/?s=Android+Tablet+Hdmi In this economic conditions it is vital to get the most you can for your purchasing dollars. So there is certainly no good reason to over pay for Android Tablet Hdmi ...
3	23	0.344	Mid With External Keyboard -Mid With External Keyboard ...	http://www.alibaba.com/showroom/mid-with-external-keyboard.html 2012 low price tablet pc hdmi, 7inch display, wifi, support external keyboard. ... android tablet external keyboard 1.Wireless keyboard for galaxy tab 2.
4	120	0.321	Tablet with External Keyboard Host - reviews topdirsearch.com	http://www.topdirsearch.com/html/Tablet+with+External+Keyboard+Host Issue 1147: RFE: Support for external keyboard, mouse: 721 people starred this issue. Comments by non-members will not trigger notification emails to users who ...
5	381	0.271	** KOCASO MID-SX9700 9.7inch Android 4.0 16GB 1.2GHz 1080p ...	http://successstory.16mb.com/kocaso-mid-sx9700-9-7inch-android-4-0-16gb-1-2ghz-1080p-hdmi-output-3d-games-wifi-ram-ddr3-1gb-capacitive-multi-touchscreen-tablet-pc-w-dual-camera-white.html KOCASO MID-SX9700 9.7inch Android 4.0 16GB 1.2GHz 1080p HDMI Output 3D Games WiFi RAM DDR3 1GB Capacitive Multi-Touchscreen Tablet PC w/ Dual Camera
6	7	0.262	Wholesale Tablet Pc External Keyboard - Buy Tablet Pc External ...	http://www.aliexpress.com/wholesale/wholesale-tablet-pc-external-keyboard.html Wholesale Tablet Pc External Keyboard from China Tablet Pc ... 10"Flytouch5 Android 2.3 Tablet PC,1GHZ CPU 4GB/8GB GPS Camera Wifi HDMI(Keyboard optional) 3G External ...
7	249	0.237	New 10.2 Superpad3/Flytouch X220 4GB 1GHz Froyo Android 2.2 ...	http://tabletpsource.com/superpad-tablet-pc/new-10-2-superpad3flytouch-x220-4gb-1ghz-froyo-android-2-2-epad-tablet-pc-with-512mb-ram-gps-bonus-keyboardstylus-penleather-case-hdmirj45-with-priority-shipping-from-usa-seller.html 10.2 Android 2.2 Epad X220 Touch Screen 512M ram 4gb +WIFI+3G+GPS+Camera+HDMI+RJ45 Detailed Product Description Epad 10.2 MID Infortm X220 HDMI output GPS tablet pc ...

8	41	0.216	#\$1: 10 GOOGLE ANDROID 4.0 TABLET 4GB FLYTOUCH 10.1 VC882 ...	http://www.antablets.com/10-GOOGLE-ANDROID-40-TABLET-4GB-FLYTOUCH-101-VC882-EPAD-LAPTOP-WIFI-CAMERA-HDMI-Support-External-3G-Flash-101-Cortex-A8WIFI-HDMI Sale 10 GOOGLE ANDROID 4.0 TABLET 4GB FLYTOUCH 10.1 VC882 EPAD LAPTOP WIFI CAMERA HDMI Support External 3G, Flash 10.1 Cortex A8,WIFI, ... Tablet Keyboard Stand.
9	175	0.215	Android Tablet Hdmi - iPads, Tablets and eReaders	http://e-bookreaderz.com/ipads-tablets-and-ereaders/android-tablet-hdmi/ Google Android 4.0 MID WiFi HDMI 1080p G-sensor 7" Tablet Capacitive TouchScreen. iPads, Tablets & eBook Readers AGPtek 2160P Netflix 3D GAMES 12 MONTH ...
10	128	0.214	Superpad3 III 3 Gen 10.2" Tablet Pc, CPU 1ghz, Google Android ...	http://tabletpcsnow.com/superpad3-iii-3-gen-102-tablet-pc-cpu-1ghz-google-android-22-102-1024x600-tft-touchscreen-webcam-gps-hdmi-usb-wifi-512mb-ram-4gb-internal-sd-support-36-gb-4-gb-32-gb-with-external-sd Buy the brilliant Superpad3 III 3 Gen 10.2 ... Superpad3 III 3 Gen 10.2" Tablet Pc, CPU 1ghz, Google Android 2.2, 10.2" 1024x600 TFT Touchscreen, Webcam, Gps, Hdmi ...

Table 10: Keywords of test case 2

Keyword	Weight
flytouch	2.450149327
superpad	2.385461106
netflix	2.250803233
epad	2.138360852
calling	1.840918058
amazon	1.636122471
vc	1.508417958
tpc	1.433648134
ebay	1.301810454
mid	1.24504779
movie	1.209320567
posttag	1.191647702
touchscreen	1.107960735
bof	1.074176712
eof	1.052347412
coby	1.046732803
froyo	0.994116226
meta	0.901472405
android	0.89937545
skype	0.889927165

Table 11: Result of test case 3

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	192	0.368	List Of Shortest People - WEBSITE REPORTED - Social Web ...	http://websitereported.com/List_of_shortest_people List Of Shortest People, websitereported, social, web, research, tool, news, images, documents, tweets, videos, posts, comments, wiki
2	16	0.323	Fastest way to Yehliu with the shortest bus ride - Taipei ...	http://www.tripadvisor.com.au/ShowTopic-g293913-i9546-k5901473-Fastest_way_to_Yehliu_with_the_shortest_bus_ride-Taipei.html Holiday Inn Express Hotel & Suites Hollywood Walk of Fame in Los Angeles; ... 9546&g=293913&faqid=392&qid=2857 When is the best time to ... Taiwan Travel Agency ...
3	401	0.318	World's Shortest Person	http://wn.com/World%27s_Shortest_Person Nepal Teen Stands Proud As World's Shortest Man, World's Shortest Man: Nepal's Chandra Bahadur Dangi measures just 54cm tall, World's Shortest Man Chandra Bahadur ...
4	6	0.253	Cheap Flights from Los Angeles to Taipei , from \$906 Round ...	http://www.farecompare.com/flights/Los_Angeles-LAX/Taipei-TPE/market.html Let FareCompare guide you to the cheapest flights from Los Angeles to Taipei. ... The shortest flight between Los Angeles, ... Travel Guides Index. Las Vegas;
5	58	0.227	Flights to Taipei - How to Get to Taiwan - Taiwan Travel ...	http://www.taiwan-travel-experience.com/flights-to-taipei.html There are less direct flights to Taipei today than 10 years ago. At that time you had more ... If you can, avoid to travel to/from Taiwan ... Los Angeles; New ...
6	141	0.207	China Adopt Talk Forum - Shortest time to LOA (800 families ...	http://chinaadopttalk.com/forum/index.php?topic=44783.0 Author Topic: Shortest time to LOA (800 families)-not expedited (Read 889 times)
7	140	0.173	Your longest and your shortest flight - SkyscraperCity	http://www.skyscrapercity.com/showthread.php?t=553862 Infrastructure and Mobility > Airports and Aviation ... Post it here. It also counts every single flight after an intermediate landing which ... Using this website ...
8	82	0.171	Flight Time from Los Angeles , CA to Taipei , Taiwan	http://www.travelmath.com/flying-time/from/Los+Angeles,+CA/to/Taipei,+Taiwan Flying time from Los Angeles, CA to Taipei, Taiwan. ... Flight time calculator. Travel Math provides an online flight time calculator for all types of travel routes.
9	1	0.169	Time Difference between Taipei , Taiwan and Los Angeles , CA	http://www.travelmath.com/time-change/from/Taipei,+Taiwan/to/Los+Angeles,+CA What is the time change from Taipei, Taiwan to Los Angeles, CA? ... Time difference. Travel Math provides an online time zone converter for places all over the world.
10	3	0.146	Cheap Flights to Taipei , Taiwan from \$725 Total Round-trip ...	http://www.asaptickets.com/cheap-flights-to-taipei ... that I as a novice group travel planner kept making, and all the time with an ... long direct flights to Taipei from its U.S. hubs in Los Angeles, ...

Table 12: Keywords of test case 3

Keyword	Weight
option	2.807868219
taipei	2.367564294
airline	2.188893296
flights	1.755954678
amp	1.704316537
ixigo	1.606406978
pm	1.477608775
am	1.435079979
shortest	1.260172204
flight	1.223065455
taiwan	1.153874984
farecompare	1.117420528
wn	1.051539997
jfk	1.024720643
airways	1.001153991
railway	0.989712754
cm	0.980224037
beijing	0.944211044
lax	0.943484909
answer	0.942829178

Table 13: Result of test case 4

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	1	0.325	Los Gatos Chinese Restaurants - Insider Pages - Restaurant ...	http://www.insiderpages.com/s/CA/LosGatos/ChineseRestaurants?cs_category=Chinese+Restaurants 20 S Santa Cruz Ave # 204, ... "You can get Chinese food anywhere or you can go to Mandarin Gourmet and get Chinese Cuisine! ... Santa Clara, CA
2	39	0.293	1867 Mandarin Way, San Jose, CA - MLS# 81224581	http://www.movoto.com/real-estate/homes-for-sale/CA/San-Jose/1867-Mandarin-Way-100_81224581.htm 0.20 ... Add or edit a restaurant. ... top-rated agents to see San Jose market statistics and to help you get the best deal for 1867 Mandarin Way, ...
3	98	0.287	El Sobrante Restaurants Eating Places in El Sobrante, CA	http://www.magicyellow.com/category/Restaurants/El_Sobrante_CA.html Fresh Mexican Cuisine In the Bay Area. Visit Us or Order Online Now.

4	3	0.273	Deals California Rainbow Rewards	http://www.rainbowrewards.com/Merchant-Search-Results.asp?t=All Your current location is California. Not correct? Then select your state from the below list.
5	50	0.272	Top Hospitality Management Schools in Santa Clara : Programs ...	http://www.onlineeducation.net/schools/hospitality-management/CA/santa-clara It is the 1838th highest ranked school in the USA and the 168th highest in the state of California ... mile radius of Santa Clara ... within 100 miles of Santa Clara
6	6	0.259	Elkins Ranch Golf Course in Fillmore, California Rankings ...	http://www.golfcourseranking.com/courses/959/California/Fillmore/93015/Elkins_Ranch_Golf_Course.html Elkins Ranch Golf Course is located in the picturesque Santa Clara River ... 20-30 Miles , Vacation Worthy ... Mountain View Golf Course Santa Paula, California 12.78 miles ...
7	91	0.255	Top Distance Education Schools in Santa Clara : Programs ...	http://www.onlineeducation.net/schools/distance-education/CA/santa-clara ... within its city limits. Santa Clara University has a total student population of 8,846. It is the 887th highest ranked school in the USA and the 57th highest in the state of California ... 20.6572 297 8.2 miles ...
8	72	0.242	California - Academic Kids	http://www.academickids.com/encyclopedia/index.php/California An Encyclopedia article about California - Academic Kids ... Template:US state symbols California is a U.S. state located on the west coast of the United States.
9	172	0.226	purchases at farmers market in Santa Barbara, Calif., on Nov ...	http://www.seasonalchef.com/bestbuys110803.htm Vintage California Cuisine: ... with mandarin oranges leading the ... I was within sight of blackened hills for a 35 mile stretch along Highway 118 before ...
10	58	0.222	Mexico Current News and Mexico Current Events, all the ...	http://mexicotoday.org/news/culture/all/node/20576?page=4 Mexico Today Ambassador and acclaimed Mexican sailor, Galia Moss, recently inaugurated a new sailboat in France in time for her next solo trip -- which will also be ...

Table 14: Keywords of test case 4

Keyword	Weight
the	3.126388
county	1.983162447
monterey	1.954339006
of	1.52771058
mateo	1.496920082
coldwell	1.43521756
scores	1.433079425
homes	1.352774149
ca	1.27053052

school	1.249736912
and	1.196651119
gatos	1.151372117
san	1.126474552
in	1.11971123
population	1.113291454
schools	1.10320502
was	1.084197591
jose	1.057783352
click	1.052460433
california	1.014210612

Table 15: Result of test case 5

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	201	0.458	SourceForge: Artificial Intelligence - Meta-Guide.com	http://www.meta-guide.com/home/ai-engine/sourceforge-artificial-intelligence Notes: This 409 item list represents a search of SourceForge for "artificial intelligence", September 25, 2011.
2	239	0.441	Dictionary - Definition of intelligence - Webster's Online ...	http://www.websters-online-dictionary.org/definitions/intelligence Earth's largest dictionary with more than 1226 modern languages and Eve!
3	8	0.429	Natural Language - AITopics / HomePage	http://aitopics.net/NaturalLanguage AAAAI's AITopics explores Natural Language Processing to ... but NLP is a term that links back into the history of Artificial Intelligence ... The current special ...
4	29	0.416	Artificial Intelligence - Pharmaceutica l Information ...	http://www.pharmainfo.net/reviews/artificial-intelligence Artificial Intelligence, ... Warwick whose work has raised the expectations of AI research far beyond its current capabilities. ... Natural language ...
5	120	0.367	artificial intelligence - Article and Reference from OnPedia.com	http://www.onpedia.com/encyclopedia/artificial-intelligence Artificial Intelligence This article is about intelligence exhibited by manufactured systems, typically computers. For other uses of the term AI, see Ai".
6	26	0.349	Artificial Intelligence Introduction	http://ai-depot.com/Intro.html Artificial Intelligence can help us understand this process by recreating it, then potentially enabling us to enhance it beyond our current capabilities.

7	223	0.343	Sample Essay: Artificial Intelligence Essay Writing Blog	http://www.genuinewriting.com/blog/sample-essays/sample-essay-artificial-intelligence/ The introduction of computers and over 50 years of research in techniques of artificial intelligence programming have led people to believe that the dream of such ...
8	253	0.334	Full text of "Artificial intelligence and expert systems ...	http://www.archive.org/stream/artificialintell27clin/artificialintell27clin_djvu.txt Full text of "Artificial intelligence and expert systems : will they change the library?"
9	243	0.331	Artificial Intelligence - Past, Present and Future Samir ...	http://www.solutionsamir.com/2008053038/Programming/Other/Artificial-Intelligence-Past-Present-and-Future.html?fontstyle=f-larger The history of artificial Intelligence Our research into the history of Artif...
10	176	0.327	Article about "Artificial intelligence " in the English ...	http://july.fixedreference.org/en/20040724/wikipedia/Artificial_intelligence AI redirects here; for alternate uses, see Ai. Artificial intelligence, also known as machine intelligence, is defined as intelligence exhibited by anything ...

Table 16: Keywords of test case 5

Keyword	Weight
the	4.536567605
of	3.743762058
and	2.339949774
to	2.149481415
ai	2.048867019
in	1.93478086
that	1.267136837
intelligence	1.242393539
is	1.151591775
artificial	0.838850116
systems	0.683856667
strong	0.602468704
turing	0.595157367
be	0.590439197
weak	0.578555042
language	0.494494858
mccarthy	0.483124654
minsky	0.462504786
system	0.398959921
human	0.389513885

Table 17: Result of test case 6

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	10	0.331	non linear curve fitting freeware - Free Download	http://www.freedownload3.com/freeware/non_linear_curve_fitting.html Math Mechanixs 1.5.0.1 A general purpose math software program and editor for solving mathematical problems and taking notes, with scientific calculator, function ...
2	93	0.311	Non linear curve fitting Free Download - Free software ...	http://www.brothersoft.com/downloads/non-linear-curve-fitting.html Non linear curve fitting Free Download, Non linear curve fitting Software Collection Download. ... linear-model fitting functions. a) the documentation, b) the ...
3	251	0.284	Non-linear editing system definition of Non-linear editing ...	http://encyclopedia2.thefreedictionary.com/Non-linear+editing+system nonlinear video editing. Editing video in the computer. Also called "nonlinear editing" (NLE), digital nonlinear systems provide high-quality post-production editing ...
4	73	0.279	Non-linear - Top-Shareware Download - Free Software Downloads ...	http://www.top-shareware.net/nonlinear.html LAB Fit Curve Fitting Software 7 ... exponential and nonlinear functions. ... Infinity is an innovative non-linear math application that allows you use complex ...
5	195	0.255	Non Linear Regression Math@Tutor Next.com	http://math.tutornext.com/statistics/non-linear-regression.html Non Linear Regression Methods Every nonlinear regression method is assumed to follow the below given steps: For each variable given in the equation, predict an ...
6	91	0.246	Curve Fitting - Free Curve Fitting Software Download	http://curve-fitting.sharewarejunction.com/ Curve fitting, free curve fitting software ... a linear and non-linear curve fitting ... and curve fitting. NLREG fits a mathematical function whose form you ...
7	117	0.239	Download Non Linear Regression Software	http://non-linear-regression.winsite.com/ Non Linear Regression software free downloads and reviews at WinSite. Free Non Linear Regression Shareware and Freeware.
8	145	0.229	Data Fitting Basics - Erithacus Software	http://www.erithacus.com/grafit/data_fitting_basics.htm Technical Support. Get technical support information about the GraFit program
9	3	0.223	nonlinear fitting Software - Free Download nonlinear fitting ...	http://www.top4download.com/free-nonlinear-fitting/ nonlinear fitting Software - Free Download nonlinear fitting ... and general nonlinear functions. DataFitting performs true nonlinear regression analysis, ...

10	452	0.216	VIPRE Antivirus Wiki SoftOasis	http://grou.ps/techwriter/wiki/tag/VIPRE%20Antivirus Trusted By Millions and Recommended Software For Home, School, Office and Entertainment
----	-----	-------	---	---

Table 18: Keywords of test case 6

Keyword	Weight
curve	4.350994158
download	2.683877577
regression	2.486153422
price	1.711796159
shareware	1.573523656
magicplot	1.560086631
curveexpert	1.524208182
windows	1.521778698
screenshot	1.440908877
findgraph	1.327385127
the	1.216025473
fitting	1.197774922
details	1.175757873
winsite	1.172396806
and	1.154825598
info	1.143308526
lab	1.12220077
mb	1.043379444
freeware	1.042126428
tags	1.027099397

Table 19: Result of test case 7

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	27	0.542	Vegetable Moussaka Feta Recipes Yummly	http://www.yummly.com/recipes/vegetable-moussaka-feta Find Quick & Easy vegetable moussaka feta Recipes! Choose from over 32 vegetable moussaka feta recipes from sites like Epicurious and Allrecipes.
2	14	0.471	Moussaka recipe bbc - CookEatShare	http://cookeatshare.com/popular/moussaka-recipe-bbc View top rated Moussaka bbc recipes with ratings and reviews. ... easy moussaka recipe; fast moussaka recipe; ... Fix it and forget it crockpot recipes;

3	24	0.301	Learn and talk about Moussaka , Arabic words and phrases ...	http://www.digplanet.com/wiki/Moussaka Greek Moussaka Recipe - How to make traditional Greek Moussaka. ... Quick Fix: Moussaka, a taste of Greece. News & Observer. Tue, 27 Nov 2012 17:13:33 -0800.
4	125	0.299	Article Archives - Quick Cooking - BellaOnline - The Voice of ...	http://www.bellaonline.com/subjects/10726.asp 10 Minute Appetizer Recipe - Pesto Christmas Tree! Make a Cute Pesto Christmas Tree for an appetizer at your upcoming Christmas Party! 10 Minute Press-In Tart and Pie ...
5	58	0.229	Ricotta Cheese Quick Recipes Yummly	http://www.yummly.com/recipes/ricotta-cheese-quick Choose from over 55 ricotta cheese quick recipes from sites like Epicurious and Allrecipes. ... Quick Moussaka. ricotta cheese, butter ... Quick Fix Lasagna. romano ...
6	175	0.196	Hungry.net - Recipe - Recipes - Food- Meal - Meat- Easy ...	http://hungry.net/ The homemade recipes you love. ... 212. Cook the beans in a large pan of boiling salted water for 4-5 mins. Cool under cold water and put in a bowl.
7	46	0.189	Moussaka in Videos	http://www.tutorgigvideo.com/v/Moussaka Definition of moussaka (oxford ... Part 1 of 2 HowToExpo.com A wonderful Middle Eastern recipe that is quick to make and full ... World of AI problem fix!
8	41	0.188	Moussaka - Pictures, posters, news and videos on your pursuit ...	http://moussaka.purzuit.com/ Quick Fix: Moussaka, a taste of Greece. News & Observer. ... Greek Three Layer Moussaka Recipe Moussaka is a layered casserole made with layers of eggplant, ...
9	54	0.181	Tigers & Strawberries » Making Moussaka	http://www.tigersandstrawberries.com/2006/08/29/making-moussaka/ She would fix me with a stern eye, ... There are a few notes I would make about moussaka, before going into the recipe: ... not too messy, and very quick. Then, ...
10	210	0.172	Seafoods recipes - HungryMonster is Restaurants, Menus ...	http://www.hungrymonster.com/recipe/recipe-search.php?C=Seafoods&ttl=1766 Restaurants, Recipes, Dining Guides, Menus, Glossaries, pricing, maps, and recipes.

Table 20: Keywords of test case 7

Keyword	Weight
sni	1.347598437
yum	1.107341247
multivar	1.081949651
wpl	0.961224336
nicest	0.878627884
soup	0.811002585
eggplant	0.787943627
moussaka	0.774644884
greek	0.762020781

lamb	0.75906395
wonderhowto	0.755828397
comment	0.727917605
recipesbest	0.721245822
how	0.691966325
reply	0.68343206
teaspoon	0.681651278
rw	0.67989072
arcamax	0.673777541
yummy	0.661579672
displaymodefull	0.659748199

Table 21: Result of test case 8

LSI Ranking	Yahoo Ranking	LSI Similarity	Short Title	URL and Abstract
1	797	0.55	Buy Gemstones Online Order Gemstone online	http://www.buygemstonesonline.com/ Buy Gemstones Online. Save Money and buy diamond earrings and other jewelry & gemstones online.
2	28	0.428	Gemstone Information and How to Shop for Them ShopGemstones	http://www.shopgemstones.com/ Unbiased information on gemstones and guide to buying. Learn how to get a good deal. Don't buy gemstone jewelry before reading this.
3	101	0.386	Loose Gemstones Loose-Gemstones .Org	http://www.loose-gemstones.org/ Loose gemstones dealer based in Phoenix, Arizona. Custom cut loose gems, commercial cut gemstones, and cabochons of all varieties are available (602) 345-1020.
4	45	0.326	Gemstones ,Semi Precious Gems , Gem Stones ,Wholesale Gemstone ...	http://www.ganpatijewelsjaipur.com/wholesale-gemstones.htm Gemstone online: Semi precious jewelry India: ... cut stone and precious cut stone. we offer you online shopping of premium quality genuine gem stone , ...
5	60	0.325	Natural Gemstones for Sale, Buy Rare & Precious Gemstones ...	http://www.vividgemstones.com/ Vivid Gemstones is a boutique store showcasing unique and rare natural gemstones of exceptional quality. We offer outstanding gemstones representing only the very ...
6	126	0.324	gem dealers, gems , precious stones	http://www.minerant.org/dealersGEM.html directory of gem and precious stones dealers in alphabetical order: A B C D E F G H I J K L M N O P Q R S T U V W Y Z A A.A. Jewel faceted stones, jewelry, rough ...
7	2	0.307	Buy Loose Gemstones Wholesale - Shop for Precious and Semi ...	http://www.gemselect.com/ Buy Gemstones Online. Loose Semi-Precious and Precious Stones: Sapphire, Spinel, Topaz, Garnet, Tourmaline, Opal, Emerald, Ruby, Amethyst and Birthstones

8	153	0.256	Classified Ads - International Gem Society	http://www.gemsociety.org/ads.htm classified ads, gems for sale, rough gems, jewelry, gemological instruments, gemologists, lapidary equipment, gem cutting service, mineral specimens.
9	64	0.248	The Jewelry Hut Gemstone Buying Guide	http://thejewelryhut.com/html/gemstone_buying_guide.html How to buy a Gemstone. The good news is we, at The Jewelry Hut, can teach about gemstone quality. After reading our Gemstone buying guide, you will know more than the ...
10	30	0.247	Gemstone Rings Online ,Buy Quality Gemstone Rings Online from ...	http://www.alibaba.com/showroom/gemstone-rings-online.html Gemstone Rings Online, Source Gemstone Rings Online Products at Bracelets & Bangles, Stainless Steel Jewelry from Manufacturers and Suppliers around the World Who ...

Table 22: Keywords of test case 8

Keyword	Weight
gemstones	3.282257418
loose	1.850326653
precious	1.78501658
gemstone	1.666544956
beads	1.600503412
gems	1.523354058
wholesale	1.383457124
ring	1.338046169
jewelry	1.329888355
currentscenario	1.322370578
rings	1.246640459
sapphire	1.157861016
semi	1.127958288
silver	1.096788765
faceted	0.964900486
buy	0.956806204
tourmaline	0.948759889
ruby	0.947534342
supplier	0.939670763
cut	0.911594097

4.2 Result Analysis

After the results were presented to Dr. Silver, he thought the results were good especially compare to the result from previous approaches. However, it would be more conclusive if we had more users to test the application.

The results demonstrate three important benefits. First, LSI ranking has about 20% of the top 10 matches that mirror Yahoo's top 10 matches. Some of LSI's top 10 matches are those that are below 50 in Yahoo's ranking. Some are even lower than 100. For example, the 10th LSI ranking page is ranked 129th in Yahoo's ranking in test case 1. The results page is the index page of some documents related the query. The page shows the tile, the abstract, and the link of each document. This information may be very helpful to some users since it serves as a summery page for the users, which is a really good starting point for their search. While the search engine providers are ranking up some pages purposely, it also allows for other biased-free ranking pages to show up.

Second, the top LSI matches are very close to the queries, which is as expected. The result return from search engine is like the first layer filter with special ranking preference. After applying LSI, the results are rearranged in a ranking without biases, making the top matches being very close to the queries. For example, the top five ranked documents from query 1 really demonstrate that. The first document is the wiki page about agent-based model (ABM). The second document is a journal. It gives details about ABM's theory, applications, and functionalities. The third document is about an application called Anylogic 6, which is used for the development of ABM. The fourth document is a professional network listing of those people who work in the relevant key word fields. It's like social network home page. The fifth document is a faculty member's course homepage. All of the search results provide a lot of information about ABM.

Third, the top keywords found by LSI are closely related to the keywords in the queries, even though there are still some stop-words. The present keyword finding method is just one of the methods to confirm the accuracy of our application. Except for those stop-words, all other keywords are closely related to the queries. To name a few,

“eggplant”, “Greek”, “lamb”, and “soup” are some examples of the keywords from query 7, which is about a recipe for Moussaka, a Greek dish with main ingredients of eggplant and lamb.

5 Conclusion

The purpose of this project is to provide bias-free search results to the users. In this project, we use LSI to reshuffle the matches return from search engine, and from this method, the ranking of the results are different and better. However, due to limited resource and funding, we cannot deploy the application by having more users test it.

6 Future Work

The biggest problem for this application now is to get the whole content for each individual match from search engine. We don't have enough hardware to get these documents in parallel. If there is enough of hardware, we can deploy the application to a small group of users and have the application tested better.

Another change we would like to make is to query against different search engine. The application made queries against Yahoo only, and it is therefore not comprehensive as Google may provide more and better raw results. By adding more hardware and including Google search service, I believe this application will become popular and more useful.

References

Grant Gross. Senators Question If Google Has Biased Search Results. PCWorld Sep 21st 2011
http://www.pcworld.com/article/240369/senators_question_if_google_has_biased_search_results.html

Grace Nasri. Is Google's Search Manipulation Hurting Consumers?
<http://www.digitaltrends.com/web/bias-and-google-shopping/>

Todd R. Weiss. Google Being Sued in UK for Bias in Search Results. eWeek January 1st 2013
<http://www.eweek.com/search-engines/google-being-sued-in-uk-for-bias-in-search-results/>

Barbar Rosario. Latent Semantic Indexing: An overview

Garcia, E. (2006). SVD and LSI Tutorial Retrieved November 28, 2006, from
<http://www.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-2-computing-singular-values.html>
<http://www.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-2-computing-singular-values.html>

M.W. Berry, S.T. Dumais. Using Linear Algebra for Intelligent Information Retrieval

Wiki page. <http://en.wikipedia.org/wiki/Tf-idf>