A Web Search Engine

Intelligent Search Engine

Vishal Kumar Malli Gunasekaran  
Computer Science  
University of Illinois at Chicago  
Chicago Illinois USA  
vmalli2@uic.edu

**ABSTRACT**

The searching strategy that is used by search engine is very important for its effective functioning and retrieving the appropriate results. With enormous data being accumulated day to day, it is very important to label the data and make use of them efficiently so that it will be easier to retrieve the data by the search engines. Otherwise it will be a challenging task and we would get irrelevant results which might not be useful.

We have built this project specifically for the domain “University of Illinois at Chicago”, we are using the components of web crawling, web page processing, indexing and intelligent searching technique and interface for the user to interact with the search engine. To start with the searching process, we are first giving the UIC URL as an input and using BFS for web crawling technique, we are collecting all the associated links up to 3000 links. An inverted index is constructed based on the indexed information. We create a bidirectional graph of associated URLs. Then the page rank score of each URL is calculated.

For a given query, TF-IDF scores for relevant URLS are calculated. Based on the cosine similarity the relevant documents are retrieved. Then these documents are sorted in descending order based on the TF-IDF score. The “intelligence” technique that is used that we combine the page rank score of each URL and TF-IDF score to retrieve the most popular pages among the relevant documents. This intelligent search technique will give a good experience for the users by retrieve appropriate pages of their interest.

In this paper, I have discussed about the main components and technique that I have used in the search, the problems encountered while building the search engine, weighing scheme and similarity measure used, ‘intelligent’ component and detailed description of the ‘intelligent’ aspect added, I have discussed the results what I have produced with my search engine and finally I have discussed the related work and future work.

**1 Main Components of Search**

* Web crawler
* Term Frequency-Inverse Document Frequency (TF-IDF)
* Page Rank
* Cosine Similarity
* Harmonic Mean
* Inverted Index
* Beautiful Soup

**Web crawler**

A Web crawler is an Internet bot which helps in Web indexing. They crawl one page at a time through a website until all pages have been indexed. Web crawlers help in collecting information about a website and the links related to them, and also help in validating the HTML code and hyperlinks. In our project we use BFS to index and crawl the URL. We restrict the domain to ‘UIC’ in our case because we are using our search restricted only to ‘UIC’. In our case the crawler is implement using multi-threading so that we can perform the search operation faster.

Inverted Index

An inverted index is a data structure index which stores a mapping to its location from its content in a document or a set of documents.

Inverted index can be classified into two categories:

a. In a record-level inverted index each word is referenced to its document.

b. In a word-level inverted index, in addition to above feature, each position of the word is also referenced.

The following actions are performed in inverted index

* Removing of Stop Words

Stop words are most occurring and useless words in document like “I”, “the”, “we”, “is”, “an”. So, these words have to be removed because they are of no use.

* Record Document IDs: If word is already present add reference of document to index else create new entry. Add additional information like frequency of word, location of word etc.
* Stemming of Root Word: Stemming of Root word is chopping of the big words and linking it with the actual words and see if we can map them. For example, if I am searching for a word called “map”, I should get the documents which has the word “mapping”, “mapped” etc.

**Beautiful Soup**

[Beautiful Soup](http://www.crummy.com/software/BeautifulSoup/) is an HTML/XML parser for Python that can turn even invalid markup into a parse tree. It provides simple, idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work. If you give Beautiful Soup a perfectly-formed document, the parsed data structure looks just like the original document. But if there's something wrong with the document, Beautiful Soup uses heuristics to figure out a reasonable structure for the data structure.

**2 Main Challenges encountered in Building the Search Engine**

1. One of the main challenges encountered was finding an optimal approach to combine TF-IDF score with the page rank algorithm
2. There were errors in text preprocessing. So, I have to fine tune it iteratively.
3. Implementation of the page rank was challenging because many of the linked URL will not be present in the crawled data. So careful measures were taken to avoid miscalculations.

**3 Weighting Scheme**

3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF, is a numeric measure that is used to score the importance of a word in a document based on how often it appeared in that document and a given collection of documents. According to our intuition, we assume that if a word appears frequently in a document, then it should be important, and we should give that word a high score. But if a word appears in too many other documents, it’s probably not a unique identifier, therefore we should assign a lower score to that word.

The math formula for this measure:

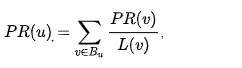
tfidf(t,d,D)=tf(t,d)×idf(t,D)tfidf(t,d,D)=tf(t,d)×idf(t,D)

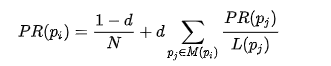
Where t denotes the terms; d denotes each document; D denotes the collection of documents.

**3.2 Page Rank**

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by Other factors like Author Rank can contribute to the importance of an entity.

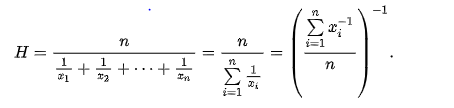
A PageRank results from a mathematical algorithm based on the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as uic.edu. The rank value indicates an importance of a page. A hyperlink to a page counts as a vote of support.





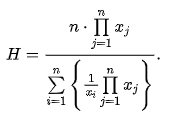
**3.3 Harmonic Mean**

The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations. The harmonic mean H of the positive real numbers is defined to be



The third formula in the above equation expresses the harmonic mean as the reciprocal of the arithmetic mean of the reciprocals.

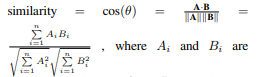
From the following formula:



**4 Similarity Measure**

Cosine Similarity

The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. In our project a calculate the total score of all the retrieved documents and store as the dictionary variables where URL is the key and the score as the values. In the descending order we sort the documents based on the cosine similarity.

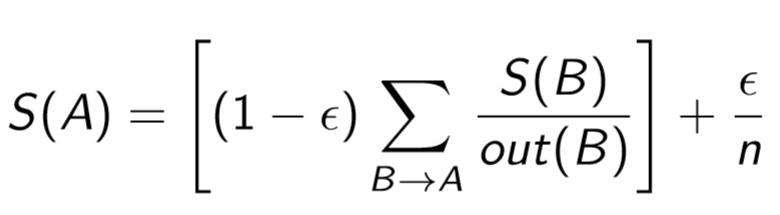
Conference Name:ACM Woodstock conference

**5 Comparison with possible alternatives**

Relevance feedback increases the weights to the web pages based on user experience, but this would require learning for the search engine to improve its efficiency

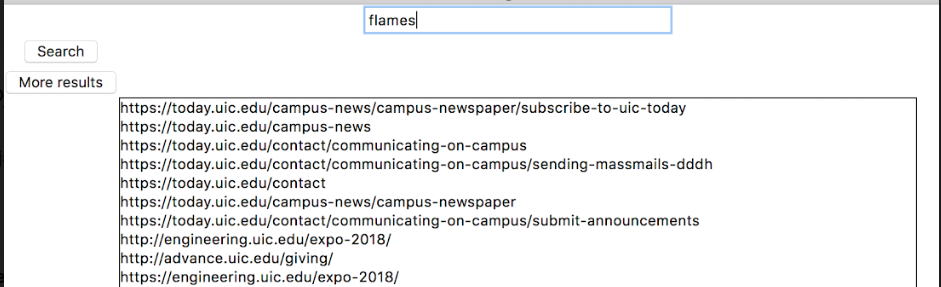
**6 Intelligent aspect added to the search engine**

The intelligent aspect that is added to the search engine is ‘page rank’ algorithm. Page rank is a link analysis algorithm and the link analysis is performed using an extended page rank algorithm to compare the result with regular query using cosine similarity. Since there are 3000 documents, page rank algorithm formula is used to run iteratively. This intelligent page rank component is integrated with cosine similarity model by taking harmonic mean of both scores.



**7. Correct Results**

a. Search word - “Flames”



TF-IDF result

('<https://today.uic.edu/category/category-flames-athletics>',  
0.4819129082426451),

('<http://flamestutoring.uic.edu/>',  
0.3197100239814187),

('<https://today.uic.edu/tag/freshmen>',  
0.3144993896271702),

('<https://today.uic.edu/tag/residence-halls>',  
0.22024694308817808),

('<https://involvement.uic.edu/>',  
0.18784542372625557),

('<http://campusprograms.uic.edu/solist.shtml>',  
0.18784542372625557),

('<http://campusprograms.uic.edu/>',  
0.18784542372625557),

('<http://www.uic.edu/depts/campusprograms/>',  
0.18784542372625557),

('[https://involvement.uic.edu](https://involvement.uic.edu/)',  
0.18784542372625557),

('<http://involvement.uic.edu/>',  
0.18784542372625557)]

Page Rank Result

('<https://www.ece.uic.edu/news/>',

0.005285924477662776),

('<https://today.uic.edu/credit-union-1-arena-celebrates-grand-opening>',0.001709920362412301),

('<http://today.uic.edu/home-away-from-home-2>',  
0.0015249287895109738),

('<http://www.uic.edu/life-at-uic>',  
0.0010687046492242186),

('<https://today.uic.edu/uic-resource-guide>',  
0.001047560767871595),

('<http://campusprograms.uic.edu/>',  
0.0010431563227592735),

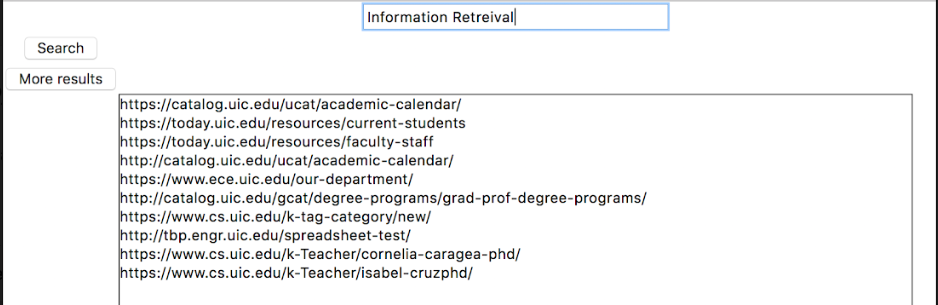
('<http://campusprograms.uic.edu/solist.shtml>',  
0.0009698594007554832),

('<https://involvement.uic.edu/>',  
0.0008342614036799586),

('[https://involvement.uic.edu](https://involvement.uic.edu/)',  
0.0007266074228439915),

('<https://today.uic.edu/credit-union-1-and-uic-poised-to-strengthen-partnership-with-naming-rights-agreement>',  
0.0007191515507281741)

b. Search word “Information Retrieval”



TF-IDF Result

[('<http://catalog.uic.edu/ucat/course-descriptions/>',  
0.17941600514511283),

('<http://catalog.uic.edu/all-course-descriptions/>',  
0.17547631982088063),

('<https://www.cs.uic.edu/k-tag-category/new/>',  
0.15830798191442505),

('<http://catalog.uic.edu/gcat/course-descriptions/>',  
0.15261479168614064),

('<http://tbp.engr.uic.edu/spreadsheet-test/>',  
0.1517941061243011),

('<https://www.cs.uic.edu/k-Teacher/isabel-cruzphd/>',  
0.1484844190539486),

('<http://catalog.uic.edu/gcat/degree-programs/grad-prof-degree-programs/>',  
0.1350021676696473),

('<https://catalog.uic.edu/gcat/degree-programs/grad-prof-degree-programs/>',  
0.1350021676696473),

('<http://catalog.uic.edu/ucat/course-descriptions/ece/>',  
0.12721623315068786),

('<https://catalog.uic.edu/gcat/course-descriptions/cs/>',  
0.12603494290897083)]

Page Rank Result

[('<https://catalog.uic.edu/ucat/academic-calendar/>',  
0.007729062641413599),

('<https://today.uic.edu/resources/current-students>',  
0.007069152204668852),

('<https://today.uic.edu/resources/faculty-staff>',  
0.006644790231563668),

('<http://catalog.uic.edu/ucat/academic-calendar/>',  
0.005132387145135511),

('<https://www.ece.uic.edu/our-department/>',  
0.004248235995792769),

('<http://catalog.uic.edu/gcat/degree-programs/grad-prof-degree-programs/>',  
0.0029445539804777592),

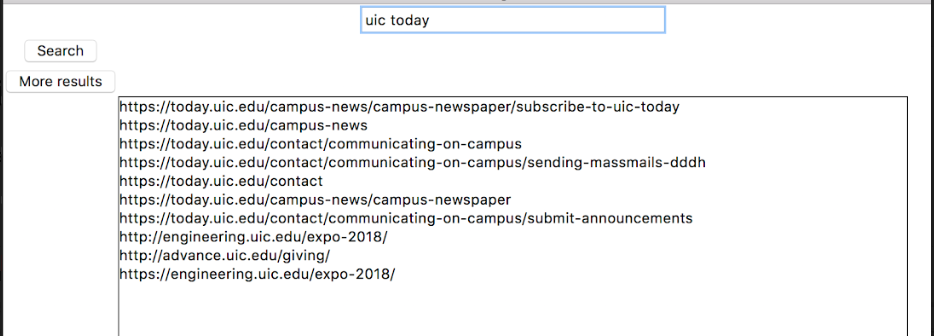
('<https://www.cs.uic.edu/k-tag-category/new/>',  
0.0009723077962278074),

('<http://tbp.engr.uic.edu/spreadsheet-test/>',  
0.0008827823516383771),

('<https://www.cs.uic.edu/k-Teacher/cornelia-caragea-phd/>',  
0.0007671694013823563),

('<https://www.cs.uic.edu/k-Teacher/isabel-cruzphd/>',  
0.0007628989257241415)]

c. Search word – “UIC today”



TF-IDF Result  
[('<https://today.uic.edu/campus-news>',

0.36752848489793033),

('<https://today.uic.edu/help-us-ignitegivingtuesday>',  
0.2414678085319482),

('<http://recreation.uic.edu/>',  
0.22079772658865515),

('[http://recreation.uic.edu](http://recreation.uic.edu/)',  
0.22079772658865515),

('<https://today.uic.edu/contact/communicating-on-campus/sending-massmails-dddh>',  
0.1618487475221226),

('<https://today.uic.edu/campus-news/campus-newspaper>',  
0.15084817082165014),

('<https://today.uic.edu/contact/communicating-on-campus>',  
0.1481048046283235),

('<http://news.uic.edu/communicating-on-campus>',  
0.1481048046283235),

('<https://www.ece.uic.edu/2017/11/09/professor-danilo-erricolo-named-university-scholar/>',  
0.11563194243497575),

('<https://today.uic.edu/campus-news/campus-newspaper/subscribe-to-uic-today>',  
0.11121013505808203)]

Page Rank Result

[('<https://today.uic.edu/campus-news/campus-newspaper/subscribe-to-uic-today>',  
0.005440189282402955),

('<https://today.uic.edu/campus-news>',  
0.004903025957402916),

('<https://today.uic.edu/contact/communicating-on-campus>',  
0.00485850024505115),

('<https://today.uic.edu/contact/communicating-on-campus/sending-massmails-dddh>',  
0.004832638083454758),

('<https://today.uic.edu/contact>',  
0.004804783592836042),

('<https://today.uic.edu/campus-news/campus-newspaper>',  
0.004757232310330939),

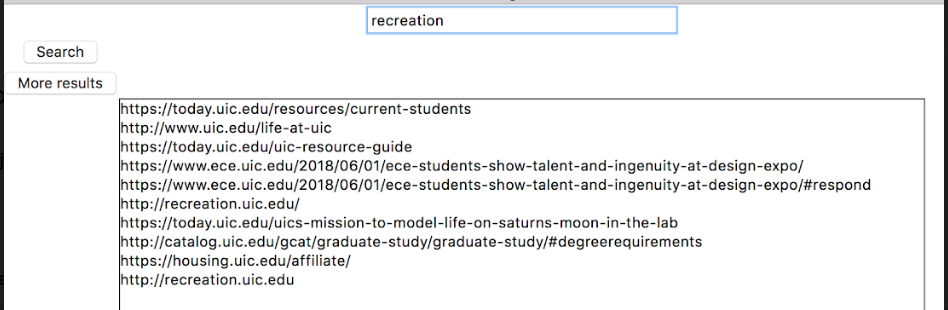
('<https://today.uic.edu/contact/communicating-on-campus/submit-announcements>',  
0.004658456821720066),

('<http://engineering.uic.edu/expo-2018/>',  
0.0019217734765484083),

('<http://advance.uic.edu/giving/>',  
0.0011545658886752924),

('<https://engineering.uic.edu/expo-2018/>',  
0.001121191721447588)]

d. Search word – “recreation”



TF-IDF Result

[('<http://recreation.uic.edu/>', 0.35298942793138044),

('[http://recreation.uic.edu](http://recreation.uic.edu/)', 0.35298942793138044),

('<http://economicimpact.uic.edu/>', 0.18531244133102914),

('<https://today.uic.edu/uic-resource-guide>', 0.15106878441405536),

('<http://www.uic.edu/life-at-uic>', 0.09872942601574762),

('<https://housing.uic.edu/affiliate/>', 0.09420777906274258),

('<https://today.uic.edu/resources/current-students>',  
0.05796213964645835),

('<https://today.uic.edu/uics-mission-to-model-life-on-saturns-moon-in-the-lab>',  
0.046070345319421446),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/>',  
0.037861039564229725),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/#respond>',  
0.037861039564229725)]

Page Rank Result

[('<https://today.uic.edu/resources/current-students>',  
0.006846571819229522),

('<http://www.uic.edu/life-at-uic>',  
0.0010663070033658862),

('<https://today.uic.edu/uic-resource-guide>',  
0.0010610860147953084),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/>',  
0.0009217277111277015),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/#respond>',  
0.0008031783910710423),

('<http://recreation.uic.edu/>',  
0.0007654231157125685),

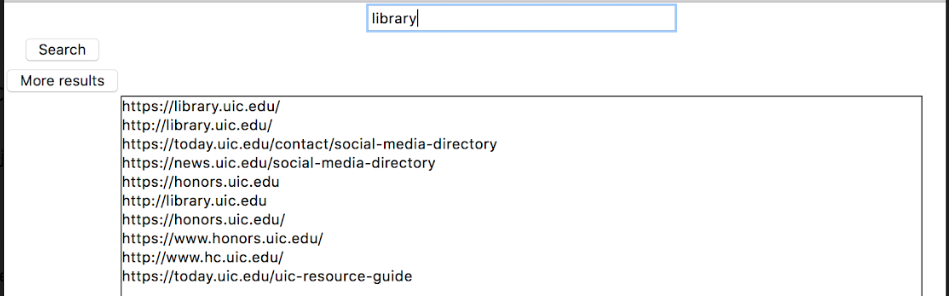
('<https://today.uic.edu/uics-mission-to-model-life-on-saturns-moon-in-the-lab>',  
0.0007467960486673603),

('<http://catalog.uic.edu/gcat/graduate-study/graduate-study/#degreerequirements>',  
0.0006635332937036185),

('<https://housing.uic.edu/affiliate/>',  
0.0006315325970597307),

('[http://recreation.uic.edu](http://recreation.uic.edu/)',  
0.00047985515545840386)]

e. Search Word – “library”



TF-IDF Result

[('<http://recreation.uic.edu/>', 0.35298942793138044),

('[http://recreation.uic.edu](http://recreation.uic.edu/)', 0.35298942793138044),

('<http://economicimpact.uic.edu/>', 0.18531244133102914),

('<https://today.uic.edu/uic-resource-guide>', 0.15106878441405536),

('<http://www.uic.edu/life-at-uic>', 0.09872942601574762),

('<https://housing.uic.edu/affiliate/>', 0.09420777906274258),

('<https://today.uic.edu/resources/current-students>',  
0.05796213964645835),

('<https://today.uic.edu/uics-mission-to-model-life-on-saturns-moon-in-the-lab>',  
0.046070345319421446),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/>',  
0.037861039564229725),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/#respond>',  
0.037861039564229725)]

Page Rank Result

[('<https://today.uic.edu/resources/current-students>',  
0.006846571819229522),

('<http://www.uic.edu/life-at-uic>',  
0.0010663070033658862),

('<https://today.uic.edu/uic-resource-guide>',  
0.0010610860147953084),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/>',  
0.0009217277111277015),

('<https://www.ece.uic.edu/2018/06/01/ece-students-show-talent-and-ingenuity-at-design-expo/#respond>',  
0.0008031783910710423), ('<http://recreation.uic.edu/>',  
0.0007654231157125685),

('<https://today.uic.edu/uics-mission-to-model-life-on-saturns-moon-in-the-lab>',  
0.0007467960486673603),

('<http://catalog.uic.edu/gcat/graduate-study/graduate-study/#degreerequirements>',  
0.0006635332937036185),

('<https://housing.uic.edu/affiliate/>',  
0.0006315325970597307),

('[http://recreation.uic.edu](http://recreation.uic.edu/)',  
0.00047985515545840386)

**8 Erroneous results**

In certain cases, for few search queries we will have result only based on page rank scores but not based on combination of TF-IDF and page rank scores because of scarcity of data.

Additionally, if there are only few search results for the query, the page rank will not work at its best.

**9 Future Work**

Combining multiple algorithms to provide better results and adopting the machine learning technique for training the results.