**Intelligent Information Retrieval**   
**CSC 575**

**Assignment 2  
Due: Thursday, February 11, 2015**

**Rong Zhuang**

1. **Retrieval from an** **Inverted Index:**  
     
   Consider the[**inverted index**](http://facweb.cs.depaul.edu/mobasher/classes/csc575/assignments/inverted-index.pdf) constrcuted from three documents (similar to the inverted index of[**Assignment 1**](http://facweb.cs.depaul.edu/mobasher/classes/csc575/assignments/assign1.html)). Using the **cosine similarity measure** determine which document is more relevant to the query: "**search engine index**". Do this by hand-tracing the retrieval algorihtm provided in slides 18 an 19 of [**Implementation Notes on Vector Space Retrieval**](http://facweb.cs.depaul.edu/mobasher/classes/csc575/lectures/implementation.pptx). Show the internediate document scores computed at each iteration. Also, show the fnal ranking and the corresponding similarity scores.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | search | engin | index | Q.Di | |Di| | SIM(Q,Di) |
| D1 | 3 | 2 | 1 | 6 | 14 | 0.9258201 |
| D2 | 0 | 0 | 2 | 2 | 4 | 0.577350269 |
| D3 | 1 | 1 | 0 | 2 | 2 | 0.816496581 |
| Q | 1 | 1 | 1 |  | 3 |  |

*\*Refer to Question1 sheet in excel.*

D1 is most relevant to the query.

1. **Indexing Models and Term Weighting**  
     
   Consider the following document-term table containing raw term frequencies. Answer the following questions, and in each case give the formulas you used to perform the necessary computations. **Note: you should not do these computations manually. You may use a spreadsheet program such as Microsoft Excel, or you can considering writing your own program do the computations., In either case, include your spreadsheet or program in your assignment submissions**.

Term1 Term2 Term3 **Term4** Term5 Term6 Term7 Term8

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

DOC1 0 3 1 0 0 2 1 0

DOC2 5 0 0 0 3 0 0 2

DOC3 3 0 4 3 4 0 0 5

**DOC4** 1 8 0 3 0 1 4 0

DOC5 0 1 0 0 0 5 4 2

DOC6 2 0 2 0 0 4 0 1

DOC7 2 5 0 3 0 1 4 2

DOC8 3 3 0 2 0 0 1 3

DOC9 0 0 3 3 3 0 0 0

DOC10 1 0 5 0 2 4 0 2

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

* 1. Compute the new weights for all the terms in document **DOC4** using the ***tf x idf*** approach.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Term1 | Term2 | Term3 | Term4 | Term5 | Term6 | Term7 | Term8 |
| DOC4 | 0.154902 | 2.40824 | 0 | 0.90309 | 0 | 0.221849 | 1.20412 | 0 |

*\*Refer to Question2 sheet in excel.*

* 1. Compute the new weights for all the terms in documents **DOC4** using the ***signal-to-noise ratio***approach.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Term1 | Term2 | Term3 | Term4 | Term5 | Term6 | Term7 | Term8 |
| DOC4 | 1.477613 | 18.04777 | 0 | 4.504189 | 0 | 1.741744 | 6.857143 | 0 |

*\*Refer to Question2 sheet in excel.*

* 1. Using the ***Keyword Discrimination*** approach, determine if **Term4** is a good index term or not (by computing it's discriminant). To compute average similarities use Cosine similarity as your similarity measure. Show your work.

Avg-sim = 0.680597, sim4 (when term 4 is removed) = 0.678385, the disc4 = 0.678385 - 0.680597= -0.00221. It’s **not** a good index term.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10 | Term1 | Term2 | Term3 | Term4 | Term5 | Term6 | Term7 | Term8 | rsv | sim |  |
| DOC1 | 0 | 3 | 1 |  | 0 | 2 | 1 | 0 | 123 | 0.741988 |  |
| DOC2 | 5 | 0 | 0 |  | 3 | 0 | 0 | 2 | 155 | 0.587458 |  |
| DOC3 | 3 | 0 | 4 |  | 4 | 0 | 0 | 5 | 244 | 0.701706 |  |
| DOC4 | 1 | 8 | 0 |  | 0 | 1 | 4 | 0 | 250 | 0.645016 |  |
| DOC5 | 0 | 1 | 0 |  | 0 | 5 | 4 | 2 | 195 | 0.671727 |  |
| DOC6 | 2 | 0 | 2 |  | 0 | 4 | 0 | 1 | 149 | 0.696231 |  |
| DOC7 | 2 | 5 | 0 |  | 0 | 1 | 4 | 2 | 241 | 0.796286 |  |
| DOC8 | 3 | 3 | 0 |  | 0 | 0 | 1 | 3 | 176 | 0.777089 |  |
| DOC9 | 0 | 0 | 3 |  | 3 | 0 | 0 | 0 | 81 | 0.446053 |  |
| DOC10 | 1 | 0 | 5 |  | 2 | 4 | 0 | 2 | 218 | 0.720292 |  |
| D\* | 17 | 20 | 15 |  | 12 | 17 | 14 | 17 | 42.80187 | 0.678385 | -0.00221 |

*\*Refer to Question2 sheet in excel.*

1. **Vector-Space Retrieval Model**

Consider the following document-term table with 10 documents and 8 terms (A through H) containing raw term frequencies. We also have a specified query, Q, with the indicated raw term weights (the bottom row in the table). Answer the following questions, and in each case give the formulas you used to perform the necessary computations. **Note:** You should do this problem using a spreadsheet program such as Microsoft Excel. Alternatively, you can write a program to perform the computations. Please include your worksheets or code in the assignment submission). [[**Download the table below as an Excel Spreadsheet**](http://facweb.cs.depaul.edu/mobasher/classes/csc575/assignments/HW2.xlsx)]

A B C D E F G H

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

DOC1 0 3 4 0 0 2 4 0

DOC2 5 5 0 0 4 0 4 3

DOC3 3 0 4 3 4 0 0 5

DOC4 0 7 0 3 2 0 4 3

DOC5 0 1 0 0 0 5 4 2

DOC6 2 0 2 0 0 4 0 1

DOC7 3 5 3 4 0 0 4 2

DOC8 0 3 0 0 0 4 4 2

DOC9 0 0 3 3 3 0 0 1

DOC10 0 5 0 0 0 4 4 2

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

Query 2 1 1 0 2 0 3 0

* 1. Compute the ranking score for each document based on each of the following query-document similarity measures (sort the documents in the decreasing order of the rank score):
     + dot product

|  |  |
| --- | --- |
| DOC2 | 35 |
| DOC7 | 26 |
| DOC4 | 23 |
| DOC1 | 19 |
| DOC3 | 18 |
| DOC10 | 17 |
| DOC8 | 15 |
| DOC5 | 13 |
| DOC9 | 9 |
| DOC6 | 6 |

* + - Cosine similarity

|  |  |
| --- | --- |
| DOC2 | 0.841726 |
| DOC7 | 0.671093 |
| DOC1 | 0.649786 |
| DOC4 | 0.565707 |
| DOC8 | 0.512989 |
| DOC10 | 0.499352 |
| DOC3 | 0.476832 |
| DOC5 | 0.439732 |
| DOC9 | 0.390199 |
| DOC6 | 0.275299 |

* + - Dice's coefficient

|  |  |
| --- | --- |
| DOC2 | 0.636364 |
| DOC1 | 0.59375 |
| DOC7 | 0.530612 |
| DOC8 | 0.46875 |
| DOC4 | 0.433962 |
| DOC10 | 0.425 |
| DOC5 | 0.4 |
| DOC3 | 0.382979 |
| DOC9 | 0.382979 |
| DOC6 | 0.272727 |

* + - Jaccard's Coefficient

|  |  |
| --- | --- |
| DOC2 | 0.466667 |
| DOC1 | 0.422222 |
| DOC7 | 0.361111 |
| DOC8 | 0.306122 |
| DOC4 | 0.277108 |
| DOC10 | 0.269841 |
| DOC5 | 0.25 |
| DOC3 | 0.236842 |
| DOC9 | 0.236842 |
| DOC6 | 0.157895 |

* 1. Compare the ranking obtained when, instead, binary term weights are used to the ranking obtained in part a where raw term weights were used (do this only with dot product as the similarity measure). Explain any discrepancy between the two rankings.

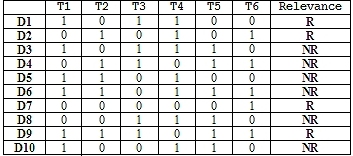
|  |  |
| --- | --- |
| DOC2 | 8 |
| DOC7 | 7 |
| DOC4 | 6 |
| DOC1 | 5 |
| DOC3 | 5 |
| DOC5 | 4 |
| DOC8 | 4 |
| DOC10 | 4 |
| DOC6 | 3 |
| DOC9 | 3 |

For binary weight ranking, the difference between values is small, it doesn’t take term frequency into account. However, dot product does consider tf, and the ranking of it is more reliable.

* 1. Construct a similar table to above, but instead of raw term frequencies compute the (non-normalized) ***tf*x*idf***weights for the terms. Then compute the ranking scores using the **Cosine similarity**. Explain any significant differences between the ranking you obtained here and the Cosine ranking from the previous part.

1. **Probabilistic Retrieval Model**

We are interested in using the following document-term matrix and the associated relevance information as training data for a probabilistic retrieval model. A 1 entry indicates that the term occurs in a document, and 0 means it does not: R or NR indicate the relevance of the document with respect to queries in the training data.



Using the basic probabilistic retrieval model, compute the relevance and non-relevance probabilities associated with terms T1 through T6 (show these probabilities in a table). Then, using these probabilities and the given query Q = (1,1,0,1,0,1), compute the discriminant   
*Disc*(Q, D11) and *Disc*(Q, D12) for each of the two new documents:

* 1. D11 = (0,1,1,0,0,1)
  2. D12 = (1,0,1,1,0,1)

Based on the discriminants, should these documents be retrieved? Explain your answer.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | T1 | T2 | T3 | T4 | T5 | T6 |
| P(R) | 0.5 | 0.5 | 0.125 | 0.5 | 0.25 | 0.75 |
| P(NR) | 0.666667 | 0.5 | 0.5 | 0.833333 | 0.833333 | 0.333333 |

*\*Refer to sheet Question4 in the excel.*

Disc(D11, Q) =  \* 

Disc(D11, Q) =  \* = 6.75

Disc(D12, Q) =  \* 

Disc(D12, Q) =  \* = 0.675

Disc(D11, Q) > 1, Q11 should be retrieved; Disc(D12, Q) < 1, Q12 should not be retrieved.

1. Read the paper [Understanding User Goals in Web Search](http://facweb.cs.depaul.edu/mobasher/classes/csc575/papers/www04--rose.pdf) by Rose and Levinson of Yahoo!. Then write a short summary (about one single-spaced page) which includes the following:
2. How can the underlying user goals in Web search be categorized and what are the primary differences between these search types?

Three main categories with sub types under them.

* Navigational

User knows clearly what he/she is searching for: a url of a website. The query should contain the name of company, university or organization, but not a famous people’s name or a film name.

* Informational

User is searching for an answer, can be open or closed. The query should contain some specific words regarding science, technology, history, news, etc. Besides, this category try to locate something in the reality world.

* Resource

Four sub categories, download, entertainment, interact and obtain. A download page for songs, videos, ebooks, software installation files; celebrities’ activity, match result, weather forecast, tax form.

1. What are some of the behavioral clues from which the search engine can deduce a user's search goals?

* the query itself
* the results returned by the search engine
* results clicked on by the user
* further searches or other actions by the user.

1. What were some of the main findings of this study and how might they be used to improve future Web search engines?
2. An issue that no approach to determine whether the goal inferred by search engine is same with user’s purpose.
3. Navigational queries seems to be less than believed before.
4. Many queries are invoked by uncertain goal, especially in the cases of obtain resource category.

Search engines need to learn knowledge of user behavior, not only how user searches, but also, why. Then, they can improve the engine’s algorithm and provide better results.