20667036 CSIT6000I

Tomas Sousa Pereira **Assignment 1** October, 2019

1.

a.

In a first glance, is allowed unlimited query review because in the end the lawyers need information to defend a lawsuit. Therefore the lawyers and the paralegals were allowed to revise the query to ensure highly effective retrieval. To support this, when made revised requests the results infered didn't have statistically significance. The revision of the query didn't change the results significantly.

b.

(i) True. The results showed us that the difference between two the lawyers was statistically insignificant. The study concluded that the results were independent of the user.

(ii) False. To proof that the amount of times the request was revised didn't change significantly the results. Blair and Maron allowed that if a lawyer wasn't satisfied with the results he could change the request. In the end the results weren't significantly to prove that if a request is revised it would get better results.

(iii) False. To reject this possibility the researchers compared the retrieval effectiveness of the lawyer against the paralegal on the same request. In some cases the lawyer got better results than paralegal, but in the end it wasn't significantly better.

(iv) False. Using the unretrieved documents were created subsets with documents which believe to be relevant documents. Then to evaluate the recall, random samples were taken and then examined by the lawyers.

(v) False. Some problems were solved by today's search engines, but there are still some problems. One of the biggest problems, but still a lot better than in that time, is that from user to user the query terms are different based on his perspective (e.g. the accident example on the study).

c.

The main reason to get low recall is due to the fact that sometimes what the users write in the query is not exactly what it is in the document, but the subject of the document. For instance sometimes the user would search "accident" and the relevant unretrieved documents only had words like "event", "situation". This happened because the manner which a person referred to the incident depends on its point of view. In the end we only wanted to retrieve documents within the subject of accidents and not only the ones with the word "accidents" in it.

d.

The experiment took almost 6 months and costed almost half a million (US) dollars.

2.

If the text passages are not updated i would try to group text passages with similar content, and infer keywords to designate that group. With this we could even reduce the redundancy in the documents (some documents can be the same). In this way we could obtain faster results for our query.
To implement the inverted index we created a class called Postings that keeps the following information: in which documents the word appear and its indexes, how many times that term appear through all the documents, the number of different documents that this word appear and all the tf-idf values for each document appeared. To complete this class was created a data structure to have correlation between the documents with its words. With this structures we could deal all the document processing in background and once finished we just needed to save the data into some type of cache. Then we just have to ask to the user the query to search for.