IR Project

AD Retrieval

Group Members:

Rahul Katinni, S20200010091

P.V. ChandraShekar, S20200010154

The main components of our IR Ad retrieval are:

1) Indexing component:

In our ir system we are performing positional indexing to build the dictionary for the ad data set that we are using. To achieve this we are pre-processing the documents in the dataset by removing all the stop words, removing unnecessary symbols, …etc. In the positional index we are storing the word as key and in how many documents it has occurred ie. document frequency and in which position it has occurred in each document.

2) Searching Component:

The search component in our ir system uses Vector Space Model to retrieve the documents which are the most closest to the query using Cosine Similarity of the query vector and all the doc vectors. We sort the retrieved similarities and give the user the top 10 most relevant searches.

3) Advanced Searching Component:

In our IR model we are also giving the user to search wild card queries. We are achieving this by building a secondary dictionary of permu-terms and all the words they are associated with. When the user searches for a wild card query we are expanding that query using the possible permu-term expansions which we have stored in our permu-term dictionary. After getting the expanded queries we are retrieving the documents relevant to all the queries that we get in the previous step.

4) Capturing Feedback:

In our IR model we are asking the user for explicit relevance feedback by asking him to mark each retrieved document as 1 or 0 for relvance and non-

relevance respectively, after we retrieve the most probable documents for his query using cosine similarity. Then after getting this relevant feedback we are storing it with the associated query and the next time a user posses a similar query to our system we take into consideration the relevance feedback given by previous user for similar queries and retrieve the documents in order of most relevant first and non-relevant in the bottom.

5) Assessments Components:

After we are taking the relevance feedback from the user we are using that data to calculate the precision and recall at each retrieval point. That is after each document is retrieved we are calculating the precision and recall values in a similar way we calculate 11-point interpolated precision recall values.