Project Proposal

Project 2 By Zachary McFarland

The probabilistic model is a common model used in Information Retrieval systems to identify the probability that a document, from a dataset, is relevant to a given user query. The probabilistic model uses an iterative process, during which the user provides feedback to further refine the results of a query. This process of user feedback is repeated until the user is satisfied with the results. One of the requirements of the probabilistic model is to compute the first iteration is to have a subset of documents from the dataset **R**. **R**, for the intentions of this proposal, represents the initial set of "relevant" documents that is presented to the user in the first step of feedback iteration. Furthermore, $\mathbf{R}^{i}$ represents the set of relevant documents as the iteration of the probabilistic model continues to refine $\mathbf{R}^{i+1}$.

This project will focus on identifying which process, from a set of predefined methods is optimal for computing **R**. Having an optimal **R** value will optimize the initial set of documents presented to the user for feedback, intern reducing the iterations the user has to participate in. If the goal of finding an optimal **R** are completed ahead of schedule, the project will begin to investigate the possibility of enhancing $\mathbf{R}^{i}$ in an effort to further reduce user participation in remaining iterations of feedback. The method being investigated to achieve this goal is discussed below.

Three methods of calculating **R** will be created in an attempt to find the most relevant result in the most efficient manner. The methods will be referred to as the following; vector, boolean, and inclusive. Each method's name is intended to be obvious to the IR model it implements to calculate **R**. The method used in our stretch goal of improving $\mathbf{R}^{i}$ calculation will be referred to as the mutual exclusion method, it too is intended to be obvious.

The vector method of computing **R** will utilize the IR model referred to as the vector model. Each document will be represented as a vector in t-dimensional space, with axis defined by orthonormal vectors representing the keywords in the dataset. queries will be executed by calculating similarity[1] between the document vector and the query vector. Initial hypothesis is that this method will have the best cumulative score.

The boolean method computes **R** using the boolean IR model. As in the boolean model, relevant documents are defined as any documents that either contain or don't contain, depending on the conjunctive query, the keywords specified in the query. This is computed by creating a term document matrix and identifying documents in the matrix that match the query parameters. Initial hypothesis is that his method will perform well for this use case, as it will not provide partial matching that is more difficult to identify with an algorithm than with human desired results.

[1] Similarity is calculated as the cosine of the angle between vector of the document vector angle and query vector angle.

The inclusive method will start by having **R** set to all documents in the dataset. The hypothesis is that this method will result in the worst iteration score, but will have the best relevancy score, the user will have to select each document that is not relevant to their query.

The requirements of this project stipulate that the *20 newsgroup* dataset, provided by the professor, shall be used as our dataset. This project uses a predefined set of queries **Q** that have a known set of desired results. A program will be created that will run **Q**, against the dataset, for each method defined to calculate **R**; vector, boolean, inclusive. The program will compare the number of iterations required to reach the desired result. It must be assumed that it is possible that a subset of the methods used may not be able to produce the desired result. Therefore the program will also calculate the inverse of how relevant the actual result is to the desired result. The results of the two calculations will be combined to form a method score that can be easily compared, the lower the score, the better the performance of the method. Regarding the stretch goal of improving $\mathbf{R}^i$ we will utilize the same calculations for finding relevance of the result to the desired result and number of iterations to achieve the desired result. The hypothesis is that we will see a correlation between the number of iterations and the relevancy score trending upward together.

The project will require the set of queries **Q** and the set of desired results for the queries **Q** as input and will result in a cumulative, relevancy, and iteration scores for each method used to calculate **R**. From these results, it can be inferred which method of calculating **R** performs the best overall, as well as for an individual use case. If a method has a low cumulative score it means it found a decently relevant set of documents in a small number of iterations, while a high score means that it performed poorly. For a specific use case a user may want to have very high accuracy, in which case they can use the method with the highest accuracy. Conversely if a user needs to find a document very quickly, but it doesn't have to be very precise, then they can choose a method accordingly. A conclusion for this project will be drawn based upon the method with the lowest overall score.