Relevance Feedback in Thematic Space using Genetic Algorithms

Adrian Rosebrock

UMBC

1 Representation of a Document

A document is simply a collection of sentences regarding a presumably central theme. To increase both syntactic and semantic retention and representation, we remove over-abstraction by breaking a document into sets of themes, which we shall call theme space. Theme space assumes that a document is a collection of themes; themes are created by related sentences; thus the same space for a document is a graph of similarity among sentences.

2 Classical Relevance Feedback

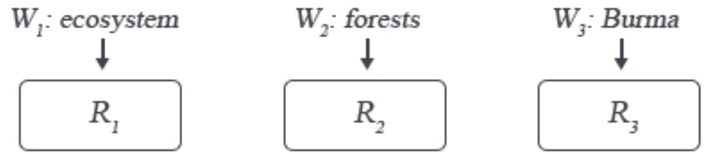
At the core, Relevance Feedback is just the refinement of a user's query to return more relevant documents. This can be done by the user selecting relevant/irrelevant documents, blindly, without the user's help, or in a psuedo-RF sense. Rochhio's Classification algorithm has long been the seminal standard.

$$\overrightarrow{Qm} = (a * \overrightarrow{Qo}) + (b * \tfrac{1}{|Dr|} * \sum_{\overrightarrow{Dj} \in Dr} \overrightarrow{Dj}) - (c * \tfrac{1}{|Dnr|} * \sum_{\overrightarrow{Dk} \in Dnr} \overrightarrow{Dk})$$

3 Problems with Classical RF

Classical RF fails to classify multimodal relationships. An example is the case of Myanmar and Burma. Both refer to the same county; however, Burma was renamed to Myanmar in 1989. Because of the representation of a document in vector space the two will be farther apart even through they are related. Theme space will attempt multimodal classification.

4 Proposed Method



After initial query, look up the themes in the resulting set of documents and construct a power set of the results for each word. Taking the intersection among elements in the power set combination, minus the null set and singleton sets, allows us to represent mutual thematic information amongst results. Each thematic intersection can be considered a centroid so we can perform normal *sim* functions. Performing feature selection on the candidate themes will return a list of tokens which will become genes for use in the genetic algorithm.

5 Evalution of System

Precision is the number of correct results divided by the number of all returned results; whereas recall is the number of correct results divided by the number of results that should have been returned. The F-measure combines both recall and precision and allows for weighting of both attributes.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

6 Future Work

Change normal Relevance Feedback to Blind Relevance Feedback and psuedo-RF. Study mutated queries from GA in hopes of simplifying the search process. Will an initial good seed to a hill-climbing algorithm be just as accurate?

7 References and Related Work

J.J. Rocchio. *Relevance feedback in information retrieval, The SMART Retrieval System : Experiments in Automatic Document Processing.* Prentice Hall Inc., 1971.

Wang Xiao-Gang and Li Yue. Relevance feedback on keyword space for interactive information retrieval. In *IITA International Conference on Services Science, Management and Engineering, 2009. SSME '09.*, 2009.