Representation of Concepts in Text: A Comparison of Within Document Frequency, Anaphora, and Synonymy

Susan Bonzi

School of Information Studies
Syracuse University
Syracuse, New York

Information retrieval systems which utilize term frequency counts are based on the assumption that the number of times a term appears in text is a good indicator of the likelihood that the concept represented by the term is central to the subject matter of the text. The simplest measure of centrality is within document frequency, where a simple count of the number of times a term or its morphological derivatives is taken. Many refinements of the measure have been introduced, for example controlling for document length and inverse document frequency, in which the collection frequencies or postings enter into the measure.

The problem with term frequency counts as a measure of "aboutness" is that the frequency with which a term occurs does not necessarily reflect the frequency of the concept represented by that term. Concepts within a text may be represented in a variety of ways: by a term, by its synonyms, and by anaphoric reference.

Synonymy is a familiar notion and is commonly used by humans when searching for information on a particular topic. The notion of anaphora is less well known in the field of information retrieval. An anaphor is a subsequent abbreviated reference to a previously specified term. For example, the phrase *a mew method for bottom watering seedlings* may be referred to later as *this method.* The word *this* refers anaphorically to the previous, fully explicated phrase.

The hypothesis of the study is that, although anaphoric resolution and mapping of synonyms to key terms will increase within document frequency, the original frequency of key terms is already higher than of terms which do not describe the document. The help that anaphoric resolution and synonym mapping render in describing the aboutness of a document is probably not worth the effort involved in refining document descriptions in these ways.

particular, related terms may be dangerous. Although recall may improve, the precision of a search might deteriorate significantly.

Not surprisingly, the length of a document does correlate with the number of anaphoric references and intellectually related terms to a particular concept. It appears from this analysis that neither anaphoric resolution nor synonym mapping is a worthwhile endeavor, however, long segments of text, e.g., full text of documents, may produce very different results.