Intelligent Techniques for Effective Information Retrieval
(A Conceptual Graph Based Approach)

Tanveer J Siddiqui
J.K. Institute of Applied Physics & Technology
Department of Electronics & Communication
University of Allahabad
Allahabad
tjs@jkinstitute.org

Abstract

With the explosive growth of information, it is becoming increasingly difficult to retrieve the relevant documents with statistical means only. This begets new challenges to IR community and motivates researchers to look for intelligent Information Retrieval (IR) systems that search and/or filter information automatically based on some higher level of understanding are required. This higher level of understanding can only be achieved through processing of text based on semantics, which is not possible by considering a document as a "bag of words". We make a humble effort in this direction by investigating techniques that attempt to utilize semantics to improve effectiveness in IR. The hypothesis is that with an improved representation of documents and by incorporating limited semantic knowledge, it is possible to improve the effectiveness of an IR system.

We propose the use of Conceptual Graph (CG) formalism for representing text. The level of semantic details to be captured within conceptual graphs and the type of relations used to relate concepts depend on the nature of the application and the domain in which it is to be used. The objective of any IR system is to classify documents as relevant and non-relevant with respect to a standing query. A detailed and accurate semantic analysis is not required for this task. This fact distinguishes IR from other natural language processing related tasks such as machine translation, question answering, text summarization, etc. Further, the amount of text with which an IR system has to deal with is so huge that it becomes impractical to carry out such analysis. So we take a middle ground by investigating techniques that make use of limited semantic knowledge that can be easily exacted from the text in the CG representation. The techniques investigated in this thesis make use of CG-based model in conjunction with the statistical vector space model. The two models thus complement each other allowing us to take the benefits of the long and established research efforts in the statistical models and versatility of semantic models.

The statistical model used in the experimental work is the most widely followed vector space model. We investigate performance of this model with different combinations of term weighting schemes for documents and queries, yielding seven different retrieval models, on the CACM-3204 collection. The best performing case, we achieved was with 'augmented normalized tem frequency' and 'raw term frequency with length normalization' components for document and query terms respectively, where inverse document frequency component is incorporated in both. The eleven-point average precision score (11-AvgP) score for the base model was 0.2961 (averaged over 64 queries).

To get CG-based representation of a document, we build conceptual graph for each sentence in the document. A small subset of basic relations has been used for constructing conceptual graphs. These relations have been extracted based on syntactic patterns. A manually constructed partial concept type hierarchy has been used. This restricts us from performing large-scale experimentation automatically. As

an alternative to this, we maintain a set of terms replaceable for each other and use some simple heuristics to capture structural variations. Queries undergo more detailed processing as compared to documents. The conceptual graphs obtained thus are stored in the form of triplets. The CG representation used is easily scalable and lends itself for fine-tuning easily according to the needs of many different NLP applications.

This thesis argues that conceptual graphs can be used as a precision tool. A hybrid information retrieval model is proposed to back up this argument. This model is essentially a two-stage retrieval model that first uses vector space model to quickly downsize the document collection and then uses CG-based model to do the final ranking. A manually constructed partial concept type hierarchy has been used. The experimental investigations have been made on CACM-3204, a collection called CGDOC, which is a small document collection designed by us and on the top ten results retrieved from LYCOS search engine. A significant improvement over the baseline performance has been observed for a subset of CACM queries.

Next, a query expansion technique based on relevance feedback has been presented. This technique attempts to reduce the 'query drift' through the use of conceptual graph based representation of document in the expansion process. Two relevance feedback strategies have been proposed and evaluated. Both the strategies resulted in a significant improvement over the baseline performance for a subset of CACM queries.

The motivation for the next technique evaluated by us lies in our belief that relation matching can be used to improve retrieval effectiveness. The roots of this belief lie in the fact that the semantics of terms can only be understood in the context they are being used. Relation matching provides a means to capture this context. To investigate the effectiveness of our technique we propose a retrieval model that integrates relation and keyword matching. There can be no better way to capture the relationships than the use of conceptual graphs and hence the CG based model, with certain enhancements and modification, remains an integral part of this work. A set of replaceable terms instead of type hierarchy has been used. Some simple heuristics along with the transitivity of relations have been used to capture structural variations. For identifying relational similarity between the query and the document, we propose new CG similarity measures. These measures have been used in the retrieval process along with the usual term similarity measures. The results obtained have shown a significant improvement (7.37%) in retrieval performance.

The results of this thesis demonstrate that by improving document/query representation and by incorporating more information from within the document and the query into the retrieval process, the effectiveness of the information retrieval is enhanced. The major findings can be summarizes as:

- The use of conceptual graphs as a knowledge representation formalism offers many advantages in IR. The general model, presented in this thesis, for creating conceptual graph from natural language text can be easily extended to capture more semantics if domain specific information is available.
- Two stage retrieval models can be quite effective in improving retrieval performance.
- Relation matching can be integrated successfully with existing statistical retrieval model to improve retrieval effectiveness.
- New CG similarity measures proposed in the thesis offer flexibility in defining textual units that make them useful for many different applications.
- Simple heuristics can be used to capture structural variations.
- Queries, instead of documents can be targeted for semantic interpretation.