Exploring and Measuring Dependency Trees for Information Retrieval

Chang Liu
School of Computing and Mathematics,
University of Ulster, Jordanstown
Northern Ireland, UK, BT37 0QB
c.liu@ulster.ac.uk

ABSTRACT

Natural language processing techniques are believed to hold a tremendous potential to supplement the purely quantitative methods of text information retrieval. This has led to the emergence of a large number of NLP-based IR research projects over the last few years, even though the empirical evidence to support this has often been inadequate. Most contributions of NLP to IR mainly concentrate on document representation and compound term matching strategies [2]. Researchers have noted that the simple term-based representation of document content such as vector representation is usually inadequate for accurate discrimination. The ”bag of words” representation does not invoke linguistic considerations and allow modelling of relationships between subsets of words. However, even though a variety of content indicator such as syntactic phrase have been tried and investigated for representing documents rather than single terms in IR systems, the matching strategy over those representation still cannot go beyond traditional statistical techniques that measure term co-occurrence characteristics and proximity in analyzing text structure.

In this paper, we propose a novel IR strategy (SIR) with NLP techniques involved at the syntactic level. Within SIR, documents and query representation are built on the basis of a syntactic data structure of the natural language text - the dependency tree, in which syntactic relationships between words are identified and structured in the form of a tree. In order to capture the syntactic relations between words in their hierarchical structural representation, the matching strategy in SIR upgrades from the traditional statistical techniques by introducing a similarity measure method executing on the graph representation level as the key determiner. A basic IR experiment is designed and implemented on the TREC data to evaluate if this novel IR model is feasible. Experimental results indicate that this approach has the potential to outperform the standard bag of words IR model, especially in response to syntactical structured queries.

1. PROPOSED IR MODEL – SIR

SIR aims to improve retrieval performance by making use of the syntactic structure of documents further, and exploring the matching strategy based on their hierarchical structural representation. SIR so far is composed of four main components.

Parsing: Minipar [1] is the parser to do the full-text parsing on the documents collection in SIR. It takes sentences as input and generate one or several pieces of dependency trees for each sentence. Documents and the query are represented by a set of raw dependency trees at this stage.

Pruning: The raw dependency trees generated directly from the parser need to be pruned for dropping some ”noisy” or extremely common words before being used as index units and the later matching. The accuracy of the tree comparison method will be optimized and not easily affected if trees are ”cleaned up”.

Indexing: The pruned dependency trees are used as index units for indexing documents so that SIR can be amenable to large-scale collection of documents. The main work is how to save the dependency tree on the disk in the form of a sequence with all the structural information retained. Subsequently, a new index architecture may be derived.

Matching: The most novel feature in SIR is to measure the relevance of documents and query on their tree representation level. The tree comparison method used in the overall matching strategy of SIR is All Common Embedded Trees (ACET) algorithm which use the number of all common embedded subtrees as a measure of similarity.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Retrieval models*

General Terms

Design

Keywords

Dependency Tree, Tree Matching, Information Retrieval

2. REFERENCES

[1] D. Lin. Minipar.
http://www.cs.ualberta.ca/ lindek/minipar.htm.

[2] T. Strzalkowski. *Natural Language Information Retrieval.* Springer, Germany, 1999.