Minimalistic test runs of the Eidetica indexer

Teresita Frizzarin (frizzarin@eidetica.com)
Annius Groenink (groenink@eidetica.com)

Abstract

Participating in a text retrieval conference for the first time, Eidetica has run six minimalistic tests with its **t·repository** indexer, doing as little tuning as possible, in order to evaluate its "performance baseline". Since no tuning was done, we will only discuss the general properties of our indexing software and how it was run on the CLEF topic sets for the monolingual German and Dutch tasks.

1. Background

Eidetica is a service provider of search and text mining technology on the basis of an application hosting model. We took part in the monolingual German and Dutch tracks of CLEF 2001, with the **t·repository** software – the core database and indexing software that drives Eidetica's hosting applications. These applications include web site search, newspaper archives, subject-based personal alerting services, automatic enrichment of XML data streams with subject keywords, and internet filtering.

The **t·repository** software subjected to the test, has the following primary characteristics:

- Built for speed, reliability and stability
- Native data input format is flat-record XML
- The source XML is compiled to a generalized index: a mathematically motivated set of *string lexicons* and *matrices* that represent relations between these lexicons. Among these matrices are both forward and backward term indexes.
- Record identifiers, data elements (authors, subjects, dates), terms, words and trigrams are all living in the same, unified, space. This allows for virtually umlimited text mining applications.
- On-the-fly context-free tagging: using simple UNIX shell pattern matching (for German/Dutch participle forms) and suffix co-occurrence rules (for singular/plural etc.), unknown words are automatically tagged and stemmed. Thus, the Eidetica software can operate with very minimalistic dictionaries. Only 5 different tags are used (noun, adword, verb, det, coor), and the tags are used purely to aid term extraction.
- Use of dictionaries is very limited. Support for force/kill lists for search terms is used very sparingly.
- Compounding: words consisting of two parts that also exist as words, are split in their "internal representation". A compound form with and without a space or dash is considered to be identical (air craft = air-craft = aircraft). This is especially useful for German and Dutch.
- The indexer performs various types of term extraction (tagging-based, proper name recognition, and extraction from so-called *full term only* - fields in the source data: controlled keyword entries, authors, etcetera). Extracted terms have a length between 1 and 4 words (or parts-of-word in the case of compounds).
- Because indexes take *terms* as entities, rather than single words, stop words and other irrelevant parts of text are automatically skipped.

2. Technique description

The topics and text were first converted (with minimal changes) to fit the profile of a single database, where records happen to contain either the fields TI, LE, TE and CAPTION (article type), or the fields TITLE, DESC and NARR (topic type). Small CLEF-specific term kill lists were made with terms such as "*relevant documents*" and *"information*" that hadn't previously been any threat to Eidetica mining applications. Then, the standard Eidetica **t·repository** indexer was run on the full document set without modifications to software or dictionaries. In other words, we have implemented *automatic runs*.

The indexer produces individual forward and backward indexes in the form of matrices between Document/Topic IDs and the terms appearing in the fields TI, LE, TE, CAPTION, TITLE, DESC and NARR. The vectors in these indexes are normalized, so that there sums are 1.

We then compute a linear combination of these indexes, where the DESC and TE fields are given a very high weight, and the remaining fields are blended in as "back-ups". The result is a matrix from document/topic IDs to terms. This matrix is transposed to a matrix from terms to document/topic IDs. Finally, we infer a Topic ID – Document ID matrix by composing the elementwise 4^{th} power root of both matrices, producing a Topic-Document rank that is proportional to the sum of the square roots of individual term co-occurrences.

This ID – ID rank matrix is the basis for the Topic-ID – Article-ID lists that we submitted in various forms.

3. Submitted runs

We have submitted the same three runs for both the Dutch monolingual task, and the German monolingual task. These runs are:

- EidNL2001A: use all topic fields (TITLE, DESC and NARR) and keep only results above a threshold score (0.80)
- EidNL2001B: use all topic fields, but produce the best 1000 results.
- EidNL2001C: use only the TITLE and DESC fields of the topics, and produce the best 1000 results.

- EidDE2001A-C: as their Dutch counterparts, but for German.