NexTrieve notebook paper Trec 2001

Gordon Clare and Kim Hendrikse

Brief description of nextrieve system, as it pertains to the trec runs

The nextrieve retrieval system is a combination fuzzy and exact search engine.

All document words are indexed.

The exact (word) index comprises approximate document position information, along with "type" information, such as the word is in a title. At the time of the TREC runs, word presence (including type) at the document level was also recorded, but word frequency in a document was not.

The fuzzy index includes text n-grams with "type" and originating word information, and their approximate document positions.

An "exact" search uses only the exact-word indexed information (namely word position and type information, and word-presence in document).

A "fuzzy" search uses the fuzzy-index information, and is assisted by a simultaneous exact word search. The "fuzziness" of a fuzzy search arises from the fact that not all query n-grams need be present in a hit for it to generate a winning score.

A score during searching is comprised of two parts -- a document level score and a proximity level score. The document level score is most important but simply collects word-presence-in-document information and, as such, does not vary on documents that contain the same set of search words with the same types.

The proximity level score of a document is the score given to the highest scoring region of the document containing the most search words (with most valuable types) in the smallest area. The position of this highest-scoring area is later used on winning documents to simplify preview generation.

Both levels of scoring have small multipliers in effect that increase as more search words are found in a particular document or region. Both levels also make use of the same scores applied to the originating words. These word level scores are generated from inverse frequency values in the database, augmented with word "type" information (giving an increase or decrease of the basic value). A "derived" penalty is also present, on words that have been automatically stemmed from an original query word.

Trec run parameterization

A few technical details of the parameterization of the TREC runs follows.

Four runs were submitted. Two were exact searches, and two were fuzzy.

- All runs were title-only, with stop words removed.
- All runs made use of a very simple stemming procedure (basically adding or removing a trailing 's' where necessary, and marking the modified word as "derived").
- All runs except ntvenx2 used a 45% increase in word score for words found in titles. Ntvenx2 used a 100% increase in word score, but this was only applied at the proximity level, not at the

document level.

ntvenx1

An exact search with a 45% increase in word score for words found in titles.

ntvenx2

An exact search with a 100% increase in word score for words found in titles. This word score increase was only applied at the proximity level, not at the document level. Recalling that the document level score is the more important score, this has the effect of removing any "type" bias at the document level, but still preserving it at the proximity level where it is nominally more important.

ntvfnx3

A fuzzy search with a setting of "minimal fuzzy". A 45% increase in word score for words found in titles was in effect. "Minimal fuzzy" has the effect of reducing the permitted word variation that can occur, and increasing the score degradation that is applied on the variation that does occur. Ie, same-letter n-grams from words who are more different from original query words get a correspondingly lower score.

ntvfnx4

A fuzzy search with a setting of "maximal fuzzy". A 45% increase in word score for words found in titles was in effect. "Maximal fuzzy" has the effect of increasing the permitted word variation that can occur, and decreasing the variation-difference score degradation that is applied.

Initial conclusions

The nextrieve search system is in a state of development and, as such, changes are being continuously made. The TREC run results weren't outstandingly good, but weren't outstandingly bad either. Considering the information being worked with at the time, the results are perfectly acceptable.

The state of the system at the time the TREC runs were submitted precluded the use of word frequency information at the document level (only presence was used).

Also, the "small multipliers" affecting scores as more words were present has been replaced by a simple counter that is both more effective and more efficient.