The NexTrieve Search System in TREC 2001

Gordon Clare and Kim Hendrikse

The NexTrieve Search System used in TREC

The NexTrieve search 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 indicating if the word is part of specially tagged text (such as a title or heading). At the time of the TREC runs, word presence (including type) at the document level was also recorded, but word frequency within a document was not.

The fuzzy index comprises text n-grams including "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-indexed 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 good or winning score.

A score of a hit during searching was comprised of two parts -- a "document level" score and a "proximity" score.
The document level score is most important but simply collected 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 had small multipliers in effect that increased as more search words were found in a particular document or region. Both levels also made 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.

Parameterization of TREC runs

A few technical details of the parameterization of the NexTrieve search engine for 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 trigrams 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.

CONCLUSIONS

The NexTrieve TREC results were not as good as expected. The best-scoring run of the four runs submitted was ntvenx2, with an average precision of 0.13. After some analysis of the NexTrieve search system results for TREC 2001 several points became readily apparent.

- The lack of word frequency information at the document level and the lack of a document length component in the scoring significantly harmed the results. Simply by adding a suitable document length metric, and adding the word frequency information, the mean average precision was increased by around 50%, to a current best of 0.19 for an exact word search.

- The presence or absence of a title-text-scoring-improvement makes very little difference to the TREC scores of NexTrieve. This is possibly due to the fact that there is, in fact, not a lot of information in titles.

- Having local ("proximity") information take part in the scoring doesn' t seem to significantly change the TREC results either. Also, the "small multipliers" affecting scores as more words were present has been removed.

- Pure fuzzy searching has several problems getting good TREC results. This is possibly due to the fact that, by its very nature, a fuzzy search uses less document-level information than is used by an exact-word search.

In short, for TREC the NexTrieve search engine must focus on document-level information in order to obtain good results. That being said, however, the other aspects of the NexTrieve search engine (fuzzy search, title text weighting, good proximity scoring) while not achieving high TREC scores are nevertheless valuable for other reasons.

Having a local (or "proximity") score take part in the overall score increases the "user-friendliness" of the system by having nicer previews arrive at the top of the result list. Ie, previews containing more of the search words appear first. This feature doesn' t improve TREC retrieval scores, but doesn' t harm them either.

Indexing title information is still valuable -- using NexTrieve it is possible to perform a search restricted to only title text (or subject text or whatever the type happens to be), resulting in significantly quicker searches. This feature doesn' t improve TREC retrieval scores, but doesn' t harm them either.

Fuzzy searching, although not providing good TREC results, still allows the user to perform searches that find information not otherwise obtainable by exact search methods. It should be noted that NexTrieve does not use any (language-dependant) stemming operations, and that the fuzzy search method employed by NexTrieve is language-independant.