

Abstract

This thesis is aimed at investigating interactive query expansion within the context of a relevance feedback system that uses term weighting and ranking in searching online databases that are available through online vendors. Previous evaluations of relevance feedback systems have been made in laboratory conditions and not in a real operational environment.

The research presented in this thesis followed the idea of testing probabilistic retrieval techniques in an operational environment. The overall aim of this research was to investigate the process of interactive query expansion (IQE) from various points of view including effectiveness.

The INSPEC database, on both Data-Star and ESA-IRS, was searched online using CIRT, a front-end system that allows probabilistic term weighting, ranking and relevance feedback.

The thesis is divided into three parts.

Part I of the thesis covers background information and appropriate literature reviews with special emphasis on the relevance weighting theory (Binary Independence Model), the approaches to automatic and semi-automatic query expansion, the ZOOM facility of ESA/IRS and the CIRT front-end.

Part II is comprised of three Pilot case studies. It introduces the idea of interactive query expansion and places it within the context of the weighted environment of CIRT. Each Pilot study looked at different aspects of the query expansion process by using a front-end. The Pilot studies were used to answer methodological questions and also research questions about the query expansion terms. The knowledge and experience that was gained from the Pilots was then applied to the methodology of the study proper (Part III).

Part III discusses the Experiment and the evaluation of the six ranking algorithms. The Experiment was conducted under real operational conditions using a real system, real requests, and real interaction. Emphasis was placed on the characteristics of the interaction, especially on the selection of terms for query expansion.

Data were collected from 25 searches. The data collection mechanisms included questionnaires, transaction logs, and relevance evaluations.

The results of the Experiment are presented according to their treatment of query expansion as main results and other findings in Chapter 10. The main results discuss issues that relate directly to query expansion, retrieval effectiveness, the correspondence of the online-to-offline relevance judgements, and the performance of the $w(p - q)$ ranking algorithm.

Finally, a comparative evaluation of six ranking algorithms was performed. The yardstick for the evaluation was provided by the user relevance judgements on the lists of the candidate terms for query expansion. The evaluation focused on whether there are any similarities in the performance of the algorithms and how those algorithms with similar performance treat terms.

This abstract refers only to the main conclusions drawn from the results of the Experiment:

(1) One third of the terms presented in the list of candidate terms was on average identified by the users as potentially useful for query expansion;

(2) These terms were mainly judged as either variant expression (synonyms) or alternative (related) terms to the initial query terms. However, a substantial portion of the selected terms were identified as representing new ideas.

(3) The relationship of the 5 best terms chosen by the users for query expansion to the initial query terms was: (a) 34% have no relationship or other type of correspondence with a query term;

(b) 66% of the query expansion terms have a relationship which makes the term: (b1) narrower term (70%), (b2) broader term (5%), (b3) related term (25%).

(4) The results provide some evidence for the effectiveness of interactive query expansion. The initial search produced on average 3 highly relevant documents at a precision of 34%; the query expansion search produced on average 9 further highly relevant documents at slightly higher precision.

(5) The results demonstrated the effectiveness of the $w(p-q)$ algorithm, for the ranking of terms for query expansion, within the context of the Experiment.

(6) The main results of the comparative evaluation of the six ranking algorithms, i.e. $w(p-q)$, EMIM, F4, F4modified, Porter and ZOOM, are that: (a) $w(p-q)$ and EMIM performed best; and (b) the performance between $w(p-q)$ and EMIM and between F4 and F4modified is very similar;

(7) A new ranking algorithm is proposed as the result of the evaluation of the six algorithms.

Finally, an investigation is by definition an exploratory study which generates hypotheses for future research. Recommendations and proposals for future research are given. The conclusions highlight the need for more research on weighted systems in operational environments, for a comparative evaluation of automatic vs interactive query expansion, and for user studies in searching weighted systems.