Enhancing Web Search through Query Expansion

Daniel Crabtree
Victoria University of Wellington, New Zealand

INTRODUCTION

Web search engines help users find relevant web pages by returning a result set containing the pages that best match the user's query. When the identified pages have low relevance, the query must be refined to capture the search goal more effectively. However, finding appropriate refinement terms is difficult and time consuming for users, so researchers developed query expansion approaches to identify refinement terms automatically.

There are two broad approaches to query expansion, automatic query expansion (AQE) and interactive query expansion (IQE) (Ruthven et al., 2003). AQE has no user involvement, which is simpler for the user, but limits its performance. IQE has user involvement, which is more complex for the user, but means it can tackle more problems such as ambiguous queries.

Searches fail by finding too many irrelevant pages (low precision) or by finding too few relevant pages (low recall). AQE has a long history in the field of information retrieval, where the focus has been on improving recall (Velez et al., 1997). Unfortunately, AQE often decreased precision as the terms used to expand a query often changed the query's meaning (Croft and Harper (1979) identified this effect and named it query drift). The problem is that users typically consider just the first few results (Jansen et al., 2005), which makes precision vital to web search performance. In contrast, IQE has historically balanced precision and recall, leading to an earlier uptake within web search. However, like AQE, the precision of IQE approaches needs improvement. Most recently, approaches have started to improve precision by incorporating semantic knowledge.

BACKGROUND

While AQE and IQE use distinctly different user interfaces, they use remarkably similar components. Both involve three components: the retrieval model, term weighting, and term selection. The four main retrieval models are the Boolean model, the vector space model, the probabilistic model, and the logical model (Ruthven et al., 2003). Term weighting is dependent on the retrieval model. For the Boolean model, the selected terms simply extend the query. For the vector space model, Rocchio (1971) developed the most common weighting scheme, which increases the weight of relevant terms and decreases the weight of irrelevant terms. The remaining models use richer weighting schemes.

Historically, query expansion approaches targeted specific retrieval models and focused on optimizing the model parameters and the selection of term weights. However, these issues have largely been resolved. The best approaches add a small subset of the most discriminating expansion terms. For web-scale data sets (those involving billions of pages), selecting about 25 terms performs best and selecting more terms decreases precision (Yue et al., 2005). Once selected, the terms are relatively easy to incorporate into any retrieval model and weighting scheme. Therefore, our focus lies on term selection.

The selected terms must address a variety of search problems. A well-known problem is low recall, typically caused by the vocabulary gap (Smyth, 2007), which occurs when some relevant pages do not contain the query terms. For example, a search for "data mining algorithms" may not find relevant pages such as one that discussed "decision trees", which used the term "machine learning" instead of "data mining". Fortunately, AQE and IQE approaches adequately address low recall for web search when the precision of the highest ranked pages is high. That makes precision improvement paramount and the focus of both recent research and this chapter.

Another well-known problem is query ambiguity, which hurts precision and arises when there are multiple interpretations for a query. For example, "jaguar" may refer to the car or the animal. As the user must clarify

their intent, AQE approaches cannot help refine these queries, but for simple cases, many IQE approaches work well. The trouble is distinguishing interpretations that use very similar vocabulary.

A recent TREC workshop identified several problems that affect precision, half involved the aspect coverage problem (Carmel et al., 2006) and the remainder dealt with more complex natural language issues. The aspect coverage problem occurs when pages do not adequately represent all query aspects. For example, a search for "black bear attacks" may find many irrelevant pages that describe the habitat, diet, and features of black bears, but which only mention in passing that sometimes bears attack – in this case, the result set underrepresents the attacks aspect. Interestingly, addressing the aspect coverage problem requires no user involvement and the solution provides insights into distinguishing interpretations that use similar vocabulary.

Query ambiguity and the aspect coverage problem are the major causes of low precision. Figure 1 shows four queries, "regular expressions" (an easy single-aspect query that is not ambiguous), "jaguar" (an easy single-aspect query that is ambiguous), "transportation tunnel disasters" (a hard three-aspect query that is not ambiguous), and "black bear attacks" (a hard two-aspect query that is ambiguous). Current search engines easily solve easy, non-ambiguous queries like "regular expressions" and most IQE approaches address easy, ambiguous queries like "jaguar". The hard queries like "transportation tunnel disasters" are much harder to refine, especially when they are also ambiguous like "black bear attacks".

Predicting query difficulty has been a side challenge for improving query expansion. Yom-Tov et al. (2005) proposed that AQE performance would improve by only refining queries where the result set has high precision. The problem is that users gain little benefit from marginal improvements to easy queries that already have good performance. Therefore, while predicting query difficulty is useful, the challenge is still to improve the precision of the hard queries with low initial precision. Fortunately, the most recent query expansion approaches both improve the precision of hard queries and help predict query difficulty.

MAIN FOCUS

There are a variety of approaches for finding expansion terms, each with its own advantages and disadvantages.

Relevance Feedback (IQE)

Relevance feedback methods select expansion terms based on the user's relevance judgments for a subset of the retrieved pages. Relevance feedback implementations vary according to the use of irrelevant pages and the method of ranking terms.

Normally, users explicitly identify only relevant pages; Ide (1971) suggested two approaches for de-

Figure 1. Query examples and factors affecting precision