Recasting the context in relevance feedback

Ian Ruthven
University of Glasgow
Glasgow, G12 8QQ, Scotland
<igr@dcs.gla.ac.uk>

The use of term co-occurrence information has a long history in information retrieval (IR). The aim is to exploit potential semantic relationships between terms that appear in the same documents. These are used to derive a structure either on the document collection (e.g. clustering) or on the terms (e.g. automatic thesaurus construction). An alternative approach is to use these relationships for relevance feedback.

A common technique for relevance feedback - so common that many researchers regard it as the only technique - is to add terms appearing in relevant documents to a query. An alternative approach is to use relevance assessments to reweight terms appearing in the relevant documents. Here these new weights are allied to co-occurrence information to produce improved weights for terms not appearing in relevant documents. This leads to a notion of *indirect* evidence. When a user selects a document, she is giving direct evidence for the relevance of the terms in the document. She can also be seen as giving indirect evidence for related terms within the database. The use of indirect evidence for a term allows the weighting of a term 'as-if' it had appeared in a relevant document. This is based on an implicit assumption in IR, namely that relevance feedback is not an independent process rather it is *relative* to a knowledge base.

The approach taken here is to measure the impact of a document selection upon a set of terms. There are two possibilities in following this approach. The first method is to modify both the weight given to a term *and* the strength of relationship between terms. The second is to treat the relationship between terms as fixed, reflecting a static, statistical dependency and to modify only the strength assigned to a term. This work follows the second line, weighting terms that appear in relevant documents and using the relationship between terms to calculate the weight of other *potentially* relevant terms. These are terms that do *not* appear in a document marked as relevant by the user, but that are related to the relevant terms.

In this model a probability value is attached to a term giving an estimate of its importance in describing the current stage of the search. The co-occurrence relationships are expressed as a conditional relationship. This gives the probability of one term appearing in a document given the presence of another term. The probability distribution over the terms in a collection can be regarded as the intial context. This is derived objectively, based on term occurrence information.

The context develops subjectively through relevance feedback. Although the initial context is comprised of all the terms in the collection future contexts are only a subset of this initial context. An important part of context development, then, is selecting an appropriate sub-context within which to interpret a document selection. During each cycle of relevance feedback the context, as shown in Figure 1, is comprised of the terms that appear in the relevant document and the terms that co-occur with these terms.

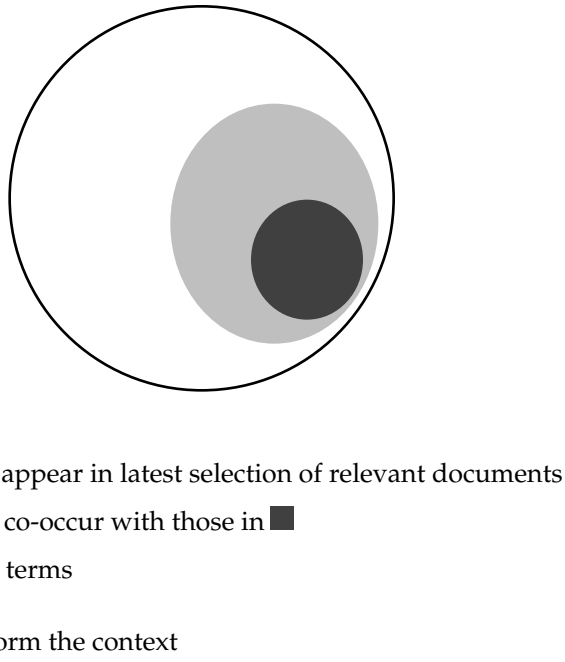


Figure 1 Context of a search

The terms in the part of the context are those for which we have direct evidence of relevance and their probability of importance can be estimated directly. The terms in the part of the context have their probability of importance assessed indirectly. The major assertion here is that, without direct evidence for a term, its importance can be inferred from the terms with which it co-occurs. Equation 1 gives the probability of $term_X$ given that it co-occurs with the set of terms C.

$$P(term_{X=1}) = \sum_{i=1}^{\#C} P(term_X = 1 \mid term_i = 1)\, P(term_i = 1) \qquad (1)$$

The conditional probability between two terms that do not co-occur is taken to be zero, that is the presence of a term does not supply any evidence for the importance of a term with which it does not co-occur. This means that equation 1 implicitly takes into account all the terms in the collection when determining the importance of a term. This mechanism implements the notion

of indirect evidence discussed above. When a user selects a document, the probability of importance of a term that appears in the document is assessed directly. The change in importance of a term is propagated to the terms with which it co-occurs. Each term that co-occurs with a relevant term is assigned a new probability using equation 1. Figure 2 shows this figuratively.

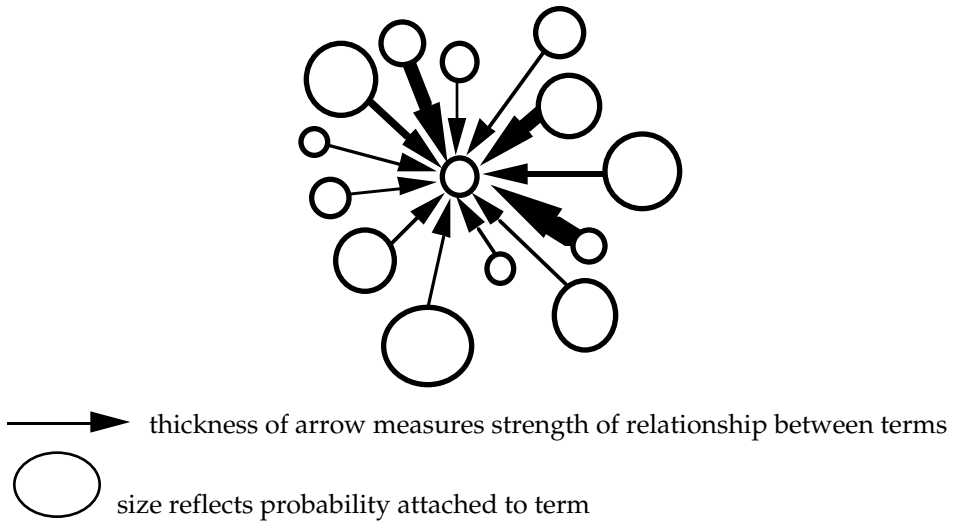


Figure 2 Calculation of probability of a potentially relevant term

Equation 2 shows how that a new probability measure can be assigned to one of these terms.

$$P'(term_X=1) = \sum_{i=1}^{\#C} P(term_X=1 \mid term_i=1)\, P'(term_i=1) \qquad (2)$$

This calculation takes into account the new probability attached to the terms in the document and the previous probability attached to the term itself. During each relevance feedback cycle the set of terms influencing the value of a term is the same. What changes in the probability values attached to these terms and to which class the terms belong.

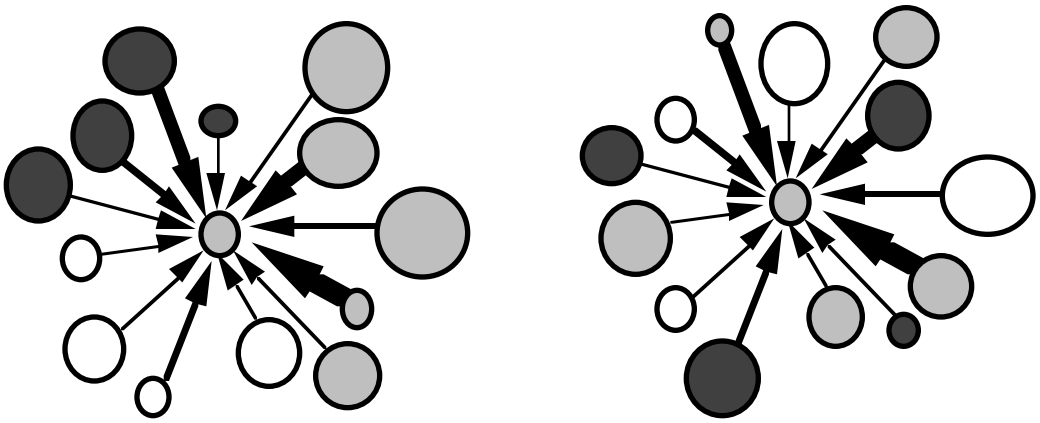
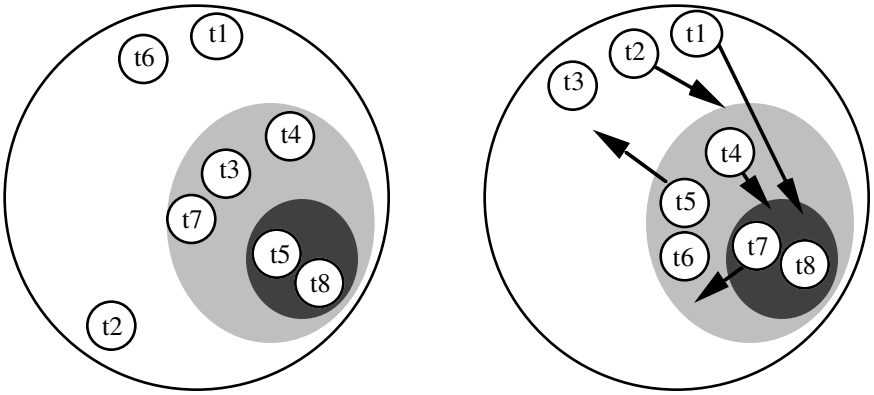


Figure 3 Change in importance and class of terms affecting the estimate of a term

Figure 4 summarises the possible changes in class of a term from one feedback cycle to another. Terms of class t1 move into the context by being present in a user-selected document, t2 move into the context by co-occurring with a term of class t1.



This system assigns a probability value to terms appearing a relevant document. It then uses these values and values of relatedness between terms to derive improved estimates for other terms in the collection. The remaining terms - those that are not 'relevant' nor co-occur with those that are, have their probabilities scaled down. This results in a new probability distribution over the set of terms in the document collection.

Each document selection(s) recasts the probability distribution. Different document selections give different contexts and different contexts lead to different retrieval results. In each relevance feedback cycle the new context is develop under document selections to provide the next context. This means that the order of document selection is important. Two users may select the same set of documents over several iterations in a search but if they select the documents in

a different order then the resulting contexts will be different. This is an attempt to capture the dynamics of an information search.