9 Relevance feedback and query expansion

In most collections, the same concept may be referred to using different SYNONYMY words. This issue, known as synonymy, has an impact on the recall of most information retrieval systems. For example, you would want a search for aircraft to match plane (but only for references to an airplane, not a woodworking plane), and for a search on thermodynamics to match references to heat in appropriate discussions. Users often attempt to address this problem themselves by manually refining a query, as was discussed in Section 1.4; in this chapter we discuss ways in which a system can help with query refinement, either fully automatically or with the user in the loop.

The methods for tackling this problem split into two major classes: global methods and local methods. Global methods are techniques for expanding or reformulating query terms independent of the query and results returned from it, so that changes in the query wording will cause the new query to match other semantically similar terms. Global methods include:

- Query expansion/reformulation with a thesaurus or WordNet (Section 9.2.2)
- Query expansion via automatic thesaurus generation (Section 9.2.3)
- Techniques like spelling correction (discussed in Chapter 3)

Local methods adjust a query relative to the documents that initially appear to match the query. The basic methods here are:

- Relevance feedback (Section 9.1)
- Pseudo relevance feedback, also known as Blind relevance feedback (Sec-
- tion 9.1.6)
- (Global) indirect relevance feedback (Section 9.1.7)

In this chapter, we will mention all of these approaches, but we will concentrate on relevance feedback, which is one of the most used and most successful approaches.

9.1 Relevance feedback and pseudo relevance feedback

RELEVANCE FEEDBACK The idea of relevance feedback (RF) is to involve the user in the retrieval process so as to improve the final result set. In particular, the user gives feedback on the relevance of documents in an initial set of results. The basic procedure is:

- The user issues a (short, simple) query.
- The system returns an initial set of retrieval results.
- The user marks some returned documents as relevant or nonrelevant.
- The system computes a better representation of the information need based
- on the user feedback.
- The system displays a revised set of retrieval results.
- Relevance feedback can go through one or more iterations of this sort. The

process exploits the idea that it may be difficult to formulate a good query when you don’t know the collection well, but it is easy to judge particular documents, and so it makes sense to engage in iterative query refinement of this sort. In such a scenario, relevance feedback can also be effective in

tracking a user's evolving information need: seeing some documents may lead users to refine their understanding of the information they are seeking. Image search provides a good example of relevance feedback. Not only is it easy to see the results at work, but this is a domain where a user can easily have difficulty formulating what they want in words, but can easily indicate relevant or nonrelevant images. After the user enters an initial query for bike on the demonstration system at: http://nayana.ece.ucsb.edu/imsearch/imsearch.html the initial results (in this case, images) are returned. In Figure 9.1 (a), the user has selected some of them as relevant. These will be used to refine the query, while other displayed results have no effect on the reformulation. Figure 9.1 (b) then shows the new top-ranked results calculated after this round of relevance feedback. Figure 9.2 shows a textual IR example where the user wishes to find out about new applications of space satellites.