User Oriented Evaluation of Information Retrieval Systems

Sung H. Myaeng
School of Information Studies
4-206 Center for Science and Technology
Syracuse University

ABSTRACT

In information retrieval, the tradition of system evaluations has been to measure effectiveness in terms of relevance of retrieved documents. While retrieval effectiveness can be measured in practice by assuming that relevance assessments of documents to queries are available from an external source to the retrieval system, the meaning of relevance can be influenced by a variety of factors such as the users' mental activities and their situations once human users become part of system operations or experiments, thereby requiring the need for more subjective, user-oriented evaluation of effectiveness. In fact, many researchers have realized testing topicality is not the same as testing relevance since topicality is not the only factor that satisfies users. In order to take into account such user- dependent factors as the purpose of using the system and comprehensibility and timeliness of documents, IR systems should be evaluated subjectively by users or information requestors, not by judges who base their decisions solely on topicality of stated queries and documents.

However, it is not always clear which aspects of "goodness" of documents are perceived to be more important than others and how they influence evaluation results. It seems necessary that more specific criteria be developed to capture different aspects of document quality and isolate one aspect from others, all of which would otherwise be intermixed and confuse the meaning of evaluation results. In our talk, we will introduce and discuss a set of "goodness" criteria, namely, fidelity, pertinence, and usefulness, as interrelated subconcepts of relevance.

Among many possible evaluation dimensions one can think of, only the three are considered based on some assumptions. Fidelity, first of all, determines how closely a document literally adheres to a stated query, regardless of user's intention. Thus, this criterion can be measured objectively by external judges as long as there exist a query and a corresponding set of retrieved documents, and is expected to result in high inter-judge agreements.

Pertinence is related to how much a document satisfies the current information need or desire for which a query is formulated. Assessments of documents on this more subjective criterion are expected to vary depending on users even if their stated queries are all the same. Retrieval effectiveness along this dimension is expected to increase as more and more user characteristics are incorporated in the retrieval process, and a better document representation is achieved with non-topical attributes of documents as well as topicality.

Usefulness is related to the user's general interests, regardless of the current information need embedded in the query. Since this dimension covers serendipitous discovery of unrequested but useful information, impertinent documents with low fidelity may still be considered useful and worth further examination as long as it satisfies the user's long-term as well as short-term interests. While pertinence judgments should be made with respect to an ideal query based on the current identifiable information need (not query), usefulness judgments are to be made with respect to the needs including those not materialized at the time of the need specification. From the system designer's point of view, this criterion needs to be taken into account in system design processes since people often browse information sources for this type of discovery and the expansion of their need.

The need for using different subconcepts in system evaluations becomes obvious when we view the notion of retrieval effectiveness in the context of Taylor's question-negotiation process with four

levels of needs, i.e., visceral, conscious, formalized, and compromised needs. We will discuss the relationship between the three criteria and the four levels of information needs at the conference.

The argument for the existence and the significance of the different goodness criteria is supported by some data derived from a series of experiments which was originally conducted in an attempt to investigate the idea of integrating user profiles into an information retrieval system. In the semi-operational retrieval situation where 10 subjects generated 30 queries, each of which was based on a real information need, judgement were made for retrieved documents from the database of Communication of ACM abstracts. The data from the experiments were analyzed to generate a hypothesis regarding the interrelationship among the three dimensions of relevance and to show why more refined criteria rather than simple relevance are necessary to evaluate retrieval systems.