Relevance

and

Information Retrieval

Kenneth Baclawski
College of Computer Science
Northeastern University

This talk is based on joint work with Bipin Indurkhya, Peter Sherwood, Dan Simovici and J. Elliott Smith.

They dined on mince, and slices of quince,
Which they ate with a runcible spoon.
– Edward Lear

Motivation

The concept of **relevance** is fundamental to modern information science and central to human communication.

The model used almost universally today in information retrieval (IR) is that relevance is a *fixed relationship* between a document and a query. The two main measures of relevance, **recall** and **precision**, presume that there is such a relationship and that it can be discovered by a panel of experts.

This use of the word “relevance” has been generally accepted since the 1930s.

Meaning of Relevance

The actual meaning of "relevance" in the everyday sense is

bearing upon the matter at hand, implying a traceable, significant, logical connection.

The everyday meaning of relevance $\neq$ the generally accepted meaning of relevance within IR.

The notion in IR is **topicality**: how well the topic of the document matches the topic of the request. In everyday speech, a topical document is "on the topic" or "about" the subject, not relevant to it.

Relevance requires a "matter at hand" for it to be meaningful. A topic without some background and intention is not sufficient to define a "matter at hand."

Early Research

The underlying model of IR has been questioned as early as the late 50s. There were intense debates about the meaning of relevance from the late 50s to the early 70s.

Two major experimental studies in the 60s showed that:

1. Relevance judgements vary from one individual to another and for the same individual over time.
2. Relevance judgements are affected by purpose, background, and levels of knowledge and skill.

These debates and experimental work have had little impact on the generally accepted rationalistic model. The same calls for more "dynamic approaches" and for a "paradigm shift" were made in the 60s and 80s.

Recent Research

A review by Schamber, Eisenberg and Nilan of the literature on relevance appeared in 1990. They go so far as to state that "the field [IR] as a whole appears to be spinning its wheels, so to speak, in terms of basic theory development."

Fortunately, there has been progress in the last decade:

1. Experimental work by several groups, including Schamber *et al.*, has shown that the cognitive state of a searcher is measurable.
2. Theoretical work by Sperber and Wilson (discourse models) and Harter is very promising.

Unfortunately, the recent work suffers from a common problem, recognized explicitly by Harter:

The authors are unable to connect their investigations with the requirements of real, operational information systems.

Proposed Framework

We propose a **framework** for IR that functions in a manner that more closely approximates the everyday sense of relevance. The framework is compatible with experimental and theoretical investigations, while at the same time addressing the need for incorporating these ideas in real systems.

The framework defines a notion of **information need** and proposes a **mechanism** whereby an IR system can understand and process information needs.

While the proposed mechanism does not address every concern raised by the literature on relevance, it does have the advantage of being based on established techniques.

Example of Information Need

The following example is due to Harter:

I am a faculty member in a school of library and information science. Among other interests I teach and do research in online searching and information retrieval. I am interested in research dealing with online searching of bibliographic databases. Specifically, I am am interested in the dynamics of the search process. I would like to learn more about any empirical research that offers insight into how people do, online searches as well as theoretical models (e.g., cognitive, probabilistic, decision-theoretic, algorithmic, or other models) related to the online search process that have been tested with empirical data.

Information Need

The proposed framework assumes that there is a **searcher** who wants to retrieve relevant documents from a corpus.

The searcher's motivation or "matter at hand" is approximated by a specification called the **information need**. The information need dynamically changes as the search proceeds.

An information need has three components:

1. **Background.** This includes the searcher's educational and disciplinary background as well as a current background in the topic area.
2. **Topic.** This is the traditional IR query.
3. **Intention.** This is the "direction" in which the searcher would like to go. It is a local goal in the search process.

Example Continued

Harter's example be analyzed in the proposed framework as follows:

I am a faculty member in a school of library and information science. Among other interests I teach and do research in online searching and information retrieval. **Background.**

I am interested in research dealing with online searching of bibliographic databases. **IR Query is "online search process of bibliographic database."**

Specifically, I am am interested in the dynamics of the search process. **Intention.**

I would like to learn more about any empirical research that offers insight into how people do online searches **Restates previous sentence but adds requirement that the research be "empirical."**

as well as theoretical models (e.g., cognitive, probabilistic, decision-theoretic, algorithmic, or other models) related to the online search process that have been tested with empirical data. **Broadens the search to allow "theoretical" research which has been tested "empirically."**

Background

The searcher's background may be further subdivided as follows:

A. **General Background.** This is part of the searcher's background with which the searcher identifies. For example, "I am a molecular biologist." The general background defines an subject area, a subdivision of the subject area and a set of documents which may be assumed to be known to the searcher.

B. **Level.** The level of the searcher ranges from *novice* to *expert.* It further refines the set of documents that may be assumed to be known.

C. **Current Background.** This is the set of documents that the individual searcher has seen. It dynamically changes during the search process. Documents in this set may be annotated with attributes such as whether the document was relevant in the recent past.

Example Continued

Continuing the analysis of Harter's example:

I am a faculty member **Level of Expertise.**

in a school of library and information science. **General Background.**

Among other interests I teach and do research **Level of Expertise.**

in online searching and information retrieval. **General Background.**

Topic of Interest

The searcher's **topic of interest** will usually be called the **query**.

An **elementary query** is a topic specification using a keyword or a knowledge structure.

Elementary queries can be combined to form **general queries** using

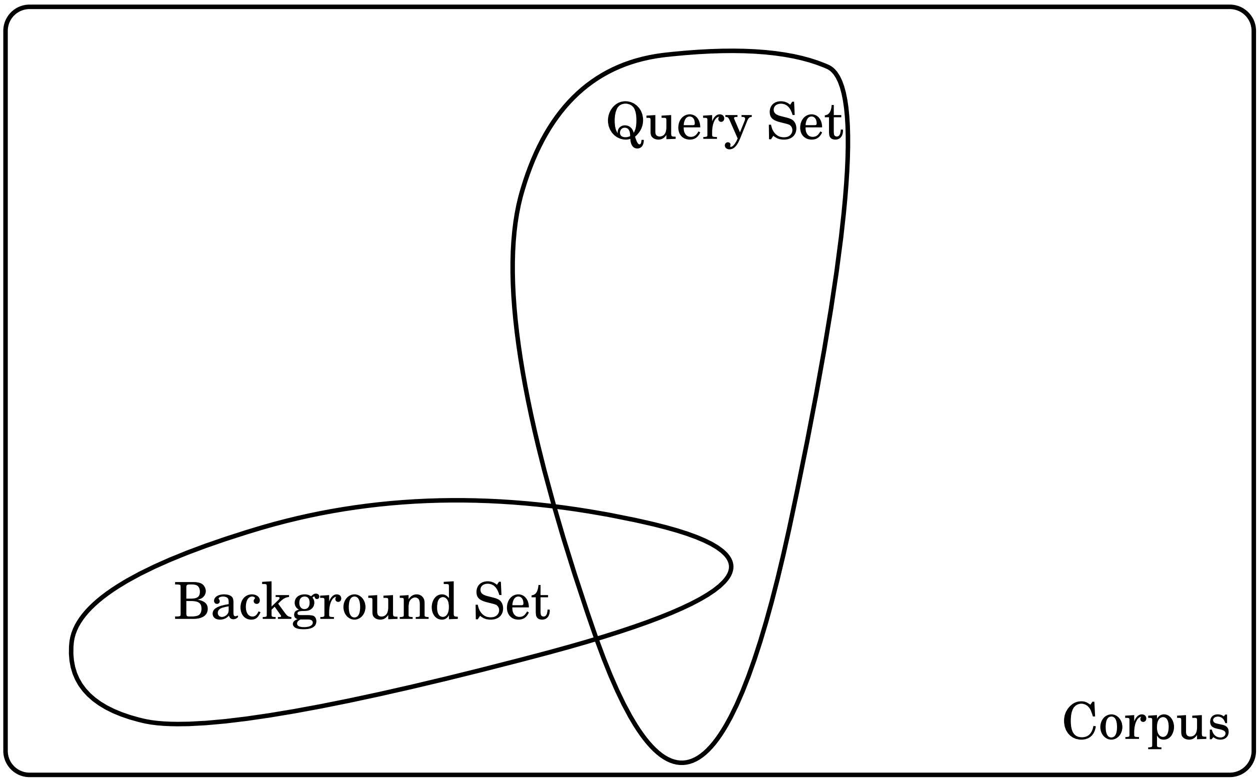
1. Boolean operators (*and*, *or*),
2. Vector space operators (linear combination) and
3. Expansion (using a thesaurus or glossary).

Implicit in any query are the topics implied by the general background and level of the searcher. As a result, an information need can include a nontrivial query even if no explicit query is specified.

Background and Query as Sets

Within a corpus of documents, the background and query define two (fuzzy) sets:

1. The Background Set. These are the documents which may be regarded as familiar to the searcher.
2. The Query Set. These are the documents that are about the same topic as the query, including both explicit and implicit topic specifications.



Phenomenological Relevance

Sperber and Wilson define **relevance of a phenomenon** by the following two conditions:

1. A phenomenon is relevant to an individual to the extent that the contextual effects achieved when it is optimally processed are large.
2. A phenomenon is relevant to an individual to the extent that the effort required to process it optimally is small.

In other words, maximize the new information (contextual effects) while minimizing the effort to process the new information.

Existing Techniques

Current IR systems are concerned almost exclusively with topicality. For a query q, let $\mathrm{top}_q(d)$ be the measure of the topicality of a document d.

Many IR systems are also capable of measuring document similarity. Let $\mathrm{sim}(d_1, d_2)$ be the measure of the similarity of documents d_1 and d_2. Such a measure is used for categorization and is not generally available to the end user.

The proposed framework requires a slightly more general measure $\mathrm{sim}_q(d_1, d_2)$ which computes document similarity relative to a topic specification q.

Document Relevance

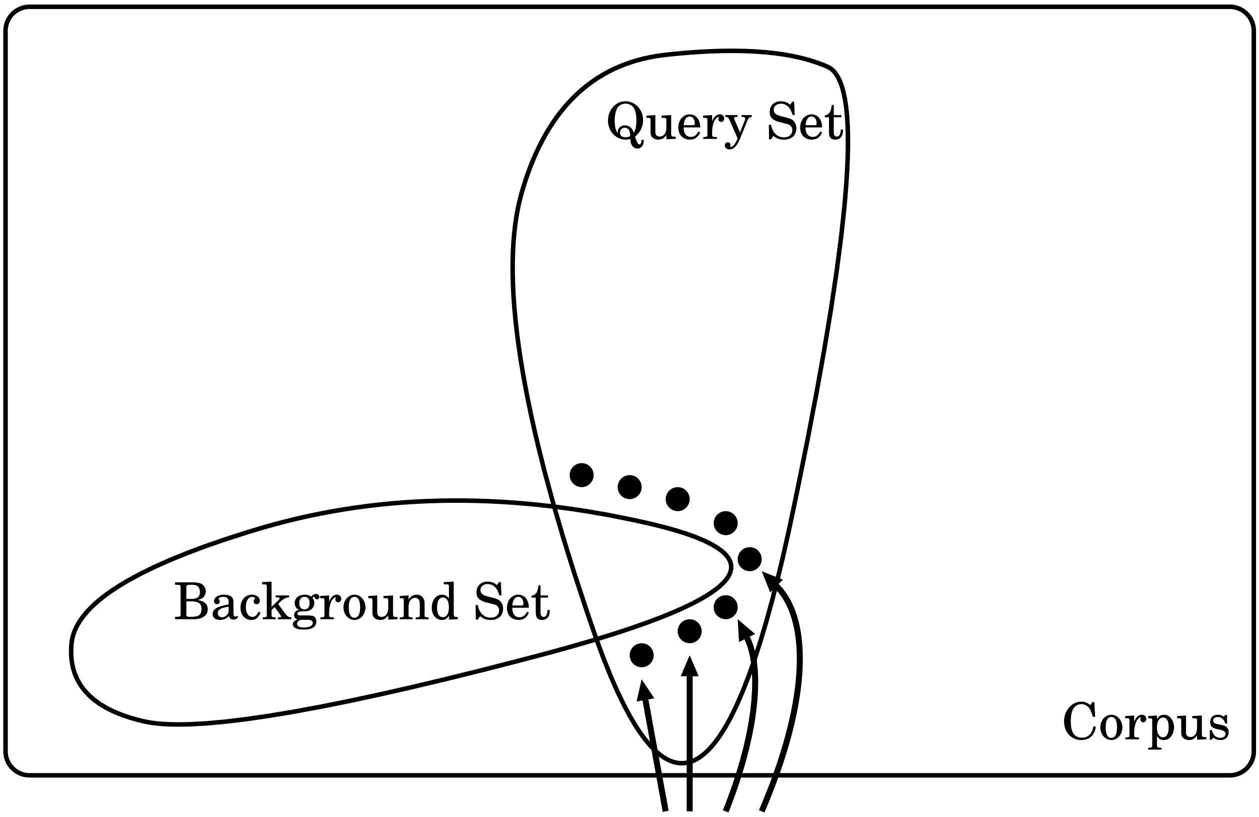
In terms of the proposed framework the Sperber-Wilson criteria may be interpreted as follows:

1. A document is relevant to the extent that it maximizes topicality but is not yet known to the searcher.
2. A document is relevant to the extent that it is close to the searcher's background.

The criteria above have to be modified when the searcher specifies an intention. General goals can be very complex, but local search intentions are often relatively simple. Intentions can lead to the paradoxical situation of a document being relevant even though it is not topical (or more precisely not topical with respect to the explicitly specified query).

Undirected Searching

The criteria for relevance can be illustrated as follows:



Documents on the boundary of the background set and also within the query set.

Specifying Intentions

In the proposed framework intentions are specified using standard database query techniques.

For this to work, one must assume that the documents have been expressed using some **knowledge representation**, such as knowledge frames or semantic networks.

It also presumes that the corpus is confined to a single subject area for which an **ontology** has been developed.

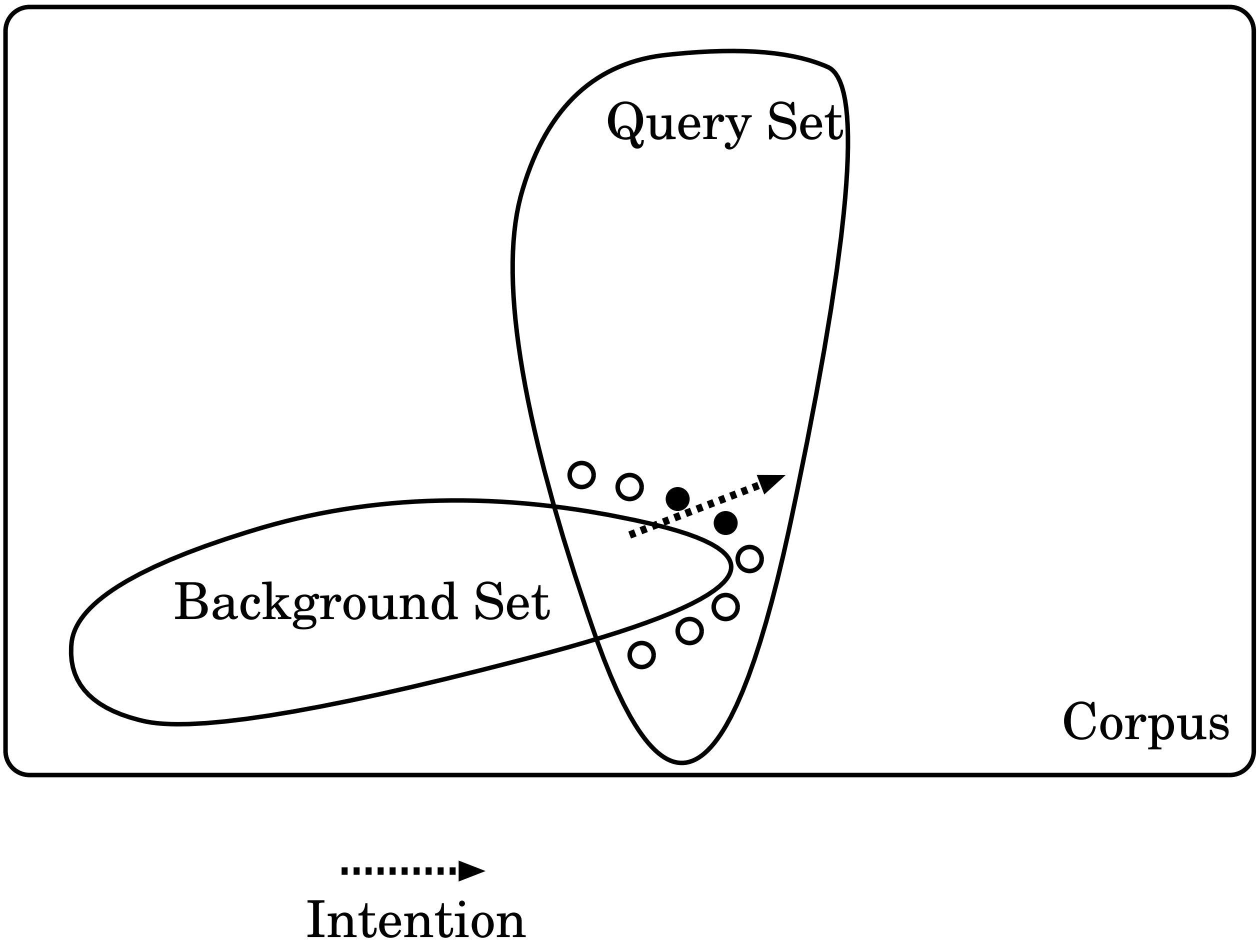
Examples of Search Intentions

The following are examples of intention specifications:

1. **Projection.** The intention is to find how a slot can be filled. In Harter's example of an information need, his intention is to fill in the "dynamics" slot of the "search process" frame.
2. **Join.** Also called inferencing. A desired frame is linked to a selected frame by a common slot value. Turtle and Croft make inferencing the basis for their approach to document retrieval.
3. **Aggregation.** The slot values for a set of selected frames are aggregated using functions like *average*, *maximum*, etc.

Directed Searching

One can illustrate the distinction between directed and undirected searching as follows:



Note that this picture is somewhat unrealistic since it is in two dimensions whereas the actual search space is very high-dimensional.

Search Strategies

The proposed framework is only a mechanism, so it allows for a variety of search strategies by manipulating the various components of an information need. Whether a particular interface would allow such flexibility is a **policy** issue.

Here are some examples:

1. **Expert.** Harter's example illustrates a fully specified information need, using all the components.
2. **Novice.** A typical novice would have neither a significant background nor a specifiable search intention. Such a searcher would be given elementary/introductory documents in the corpus.

Search Strategies Continued

3. **Sophisticated Novice.** This is a searcher who is sophisticated as a searcher, but a novice within a particular subject area. Such a searcher would have no significant background in the subject area but could have very well specified search intentions.
4. **Patent Search.** The proposed patent is given as the background, and no search intention is specified.
5. **Reviews.** Many documents simply summarize or review other documents. A summarization can be obtained by specifying a topic, a minimal background and an aggregation search intention. For example, "What are the most commonly grown genotypes/phenotypes of *E. coli*?"

Conclusion

A new framework for IR has been proposed. The new framework more closely approximates the everyday meaning of relevance.

A mechanism has been proposed whereby a searcher can express an information need, and an IR system can determine the most relevant documents in this context. The proposed mechanism:

1. Combines boolean, vector space and thesaurus query mechanisms;
2. Combines the traditional IR concept of a query (fuzzy and topic-based) with the DB concept of a query (precise; using selection, projection, join and aggregation); and
3. The framework uses existing techniques for retrieval: no new measures must be invented.