Why a corpus-topics-relevance judgments framework isn't enough: two simple retrieval challenges from the field

Catherine C. Marshall

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA

cathymar@microsoft.com

ABSTRACT

In this paper, I use two challenges to illustrate how retrieval tasks can fall outside the current corpus-topics-relevance judgment evaluation framework. The two challenges that span both desktop search and standard information retrieval are: (1) encountering new information or re-encountering forgotten information and (2) retrieval of the appropriate version of semi-redundant information, either from a personal information space or from a public store of datasets. These challenges probe the assumptions underlying corpus construction, topic selection, and relevance judgment by suggesting some common activities violate them.

1. INTRODUCTION

Information retrieval researchers have developed a strong sense of community and an equally strong sense of the value of specific research contributions by focusing on a core set of metrics, standard evaluation methods, and reference corpora, queries, and relevance judgments. It is easy to measure progress against this backdrop – how different retrieval algorithms trade off against one another, whether tuning a particular algorithm produces better results, and how new strategies for filtering and retrieval measure up to old ones. Thusfar, this agreed-upon evaluation backdrop has paid off: the community has made significant progress on a variety of information retrieval problems, and by applying analogous evaluation techniques, has been able to assess progress in new areas (such as question answering) and retrieval methods that address new media types (for example, video).

We would thus expect challenges to the efficacy of these evaluation metrics and methods to arise when a search activity or corpus characteristics are at odds with the community's assumptions about information and how it is used. Of course, the fluidity, variability, and distribution of the information on the Web and the enormous range of information needs of the people who search it – not to mention the adversarial nature of Web search – pose a striking challenge to information retrieval business as usual; these challenges have been reflected in the addition of new TREC tracks as well as in changes to the strategies used by commercial search engines. But it is still helpful to identify specific types of situations in which the evaluation criteria do not quite apply.

In this workshop paper I focus on two such challenges, both of which seem to be characteristic of self-managed information spaces in addition to external information spaces such as corpora, digital libraries, or the Web. One such challenge arises from retrieving material from what we might think of as "messy" information spaces, informal collections with semi-redundant items, such as multiple versions of the same document, perhaps with changes or revisions. The other such challenge arises from opportunistic information behavior such as clipping, in which people save or share items they have encountered in venues like magazines, newspapers, on the Web, or on broadcast media. Both of these challenges represent relatively ubiquitous everyday situations as people interact with information. Table 1 summarizes these challenges.

Table 1. Two overarching challenges

Challenges	*desktop search*	*standard IR*
Encountering new or forgotten information	Re-encounter of forgotten information in the searcher's file system or in the searcher's personal information space	Encounter of interesting, useful, or sharable information on the Web or in an archive or online publication
Retrieval of appropriate version of semi-redundant items	Locating the appropriate copy of an edited item from a personal digital store, given the searcher's information needs and expectations	Locating the appropriate copy of a revised item, given the searcher's information needs and the item's provenance

2. VERSIONS & DISTRIBUTION

Much information retrieval evaluation to-date assumes a "clean" information space that has been curated to remove items deemed to be duplicates. What happens when the information space is messy? At first blush, messiness doesn't seem like that much of a problem; there are many techniques to remove duplicate items from information spaces and it's easy to factor this kind of redundancy into evaluation. But what if the expected use of the item is central to the relevance judgment of which copy is the right one?

Field studies reveal that cleaning a collection to remove duplicates is not always as simple as it may seem; there's no straightforward heuristic that distinguishes among seemingly equivalent items. A recent field study of personal digital archiving practices revealed that consumers make copies of files as a hedge against storage catastrophes and accidental deletions [7]. This practice is not formal, but rather people welcome the opportunity to create additional versions of valued digital items. Table 2 shows the trajectory of an informant's photo of herself, one that she was very fond of. Normal use has resulted in 12 versions of a single original; the photo is now in two different formats and the jpegs are in at least two resolutions. The file also has four different names and is stored on six file systems. (I use a photo

because it's a real example; this could just as easily been a text document.)

Table 2. Tracking 12 versions of a single original photo

Description of photo file	Filename
Original on camera flash memory	126-2162_IMG.jpg
File copy on old desktop hard drive	126-2162_IMG.jpg
File copy edited in Photoshop	Eden20.psd
File copy in "sent" mail (sent to art partner who maintains web site)	Eden20.psd
File copy uploaded to web site (converted to jpeg and resolution reduced)	Eden20.jpg
File copies written to CD (as hard drive backup)	Eden20.psd & 126-2162.jpg
File copies restored from CD to new PC hard drive	Eden20.psd & 126-2162.jpg
File copy downloaded from website because psd files won't open	EB.jpg
File re-edited in photo-editing application	EB-4U.jpg
File in "sent" mail (emailed to "boys")	EB-4U.jpg

Suppose we're evaluating a very clever desktop image search algorithm. One information need she expressed (by browsing the filesystem, just to complicate matters) was to find this photo so she could attach it to an email message and send it to a prospective boyfriend she met via Match.com. Which copy of the photo counts as the right response to her query? How about the copy that's the appropriate resolution to send in an email message? How about the one that was edited in Photoshop and is now stored offline? How about the one that's the original version downloaded from the camera? Surely the nature of the task, the distribution of the data, and subtleties of the replication process should play into the presentation and evaluation of the results.

Interestingly, emerging e-science collections (especially those arising out of "little" science) yield similar types of examples [1]. For example, consider a situation in which a scientist curates a central collection of datasets. These datasets include contributions of comparable local datasets from other scientists worldwide. But each dataset is downloaded from the central site and used in many ways and many copies have been made along the way; gaps in the data are filled using different conventions and the data have been cleaned relative to different uses. For some uses, portions of the dataset are irrelevant. In other copies, measurements have been removed because the scientist using the data believes these measurements to be erroneous ("It's never 80 degrees in Greenland! This sensor must be collecting inaccurate values."). Which version is the right one? Without downloading all of them, the searcher can only tell through the use of visualization tools that run on the server side.

3. ENCOUNTERING

When we develop evaluation methods to assess algorithms, at the most basic level we assume that someone is looking for something, or – even if they are not – that they are engaged in some activity that would benefit from additional relevant information, as they would in Implicit Query scenarios [3]. But an important component of our interaction with various types of media – newspapers, broadcasts, magazines, even conversations with our friends and colleagues – is encounter [4]. Serendipitous encounter with information is apt to spur exploration, discovery, and creativity. People also use encountered information socially: they share encountered material for a variety of reasons, and although the material that they share need not be central to a current activity, it does need to have connections that are meaningful to both the sender and the recipient [6].

Recently there have been a number of efforts to explore peoples' need to re-find items they have sought in the past or encountered serendipitously (e.g. [2]); however, this is only part of the problem. As we look into the longer term relationships that people have with information – their personal archives – we find that people may not search for these things again because they don't remember that they have them [6]. This holds especially true for encountered information, because it was saved outside of an information-needs context. Yet in several different studies, when our participants re-encountered certain kinds of particularly evocative information – things they'd saved to remind them of a place or an event in their lives – they appeared to derive great pleasure from coming upon these things again. It is difficult to think of these things as falling within current information retrieval evaluation paradigms.

These challenges are not intended to suggest that the IR community abandon the current mode of competitive evaluation that conforms to an established pattern of corpus-topics-relevance judgments. Instead these challenges – and indeed the proliferation of tracks and corpora at TREC – highlight a need to examine the assumptions that underlie the evaluation strategy. Because the cases described here are seen: (1) as outside the information retrieval rubric (e.g. information that is saved without a need/encounter); (2) as human failings (e.g. forgetting what is saved in personal archives/re-encounter); (3) as information sloppiness (e.g. uncontrolled replication of personal digital data/choice among similar copies); or (4) as part of an invisible process (e.g. scientific data cleaning/redundant datasets), and found with a human in the loop [5], they have a tendency to fade from sight. Yet they are all important – and very real – examples of how people claim and reclaim information.

4. REFERENCES

[1] Borgman, C., Wallis, J., Enyedy, N. Building Digital Libraries for Scientific Data. to appear *Proc. ECDL 2006.*

[2] Bruce, H., Jones, W., Dumais, S. Information behaviour that keeps found things found. *Info. Research, 10*, 1, 2004.

[3] Cutrell, E., Dumais, S., and Teevan, J. Searching to Eliminate Personal Info. Management. *CACM, 49*, 1, 58-64.

[4] Erdelez, S. Information Encountering: A conceptual framework. In *Proc. Information Needs, Seeking, and Use in Different Contexts*. Taylor Graham, 1997, 412-421.

[5] Marchionini, G. Toward Human-Computer Information Retrieval. ASIST Bulletin, *22*, 5, June/July 2006, 20-24.

[6] Marshall, C.C. and Bly, S. Saving and Using Encountered Information. *Proc. CHI'05*, 111-120.

[7] Marshall, C.C., Bly, S., and Brun-Cottan, F. The Long Term Fate of Our Personal Digital Belongings. *Proc. Archiving 2006* (Ottawa, Canada, May 23-26, 2006) 25-30.