Web Information Retrieval

Author Preprint for Web

Nick Craswell and David Hawking

18 April 2009

1 Introduction

This chapter outlines some distinctive characteristics of web information retrieval, starting with a broad description of web data and the needs of web searching users, then working through ranking and design issues that arise. It is intended as an overview of what makes web information retrieval different and provides an introduction to some fundamental literature in the area.

2 Distinctive characteristics of the Web

The characteristics that make Web search different from some other types of search stem mainly from characteristics of the data and from distinctive web user behaviour.

2.1 Web data

Web documents are known as 'pages', each of which can be addressed by an identifier called a uniform resource locator (URL). For example: `http://www.acm.org/pubs/contents/proceedings/series/sigir/`. Web pages are usually grouped into 'sites', sets of pages published together. For example: `http://www.acm.org`

Pages are remarkably versatile. A web page can play the same roles as a news article in a conventional IR corpus, such as providing information for a report or helping to answer a question — indeed these days most news articles are also published as web pages. However, the Web also contains pages of many other types.

Many web pages exist purely to help users navigate to other pages in the site, such as the 'entry page' or 'home page' of a site. Some pages provide an interactive service such as a search form. Some pages provide a commercial service, allowing users to shop for products online. Some pages are generated dynamically and intended to be used once, such as a search engine results list

page. Other pages are front-ends to databases, so each page represents part of an underlying relational or XML database. That database might be generally useful (`http://imdb.com`) or very large and specialised (`http://www.ebi.ac.uk/flybase/` – A Database of the Drosophila Genome) to the extent that it is not clear that a general-purpose engine should index it. Some pages are error pages, indicating that the user has reached a bad URL, and others are intended to direct the user to other pages ('you are now leaving NASA' or 'doorway page to my site'). Some pages exist to contain a media file or an interactive game. Web pages are used in these ways and many more, and we often observe a single page that can be used in several of these ways. The notion of types of page, and types of user interaction, will be revisited throughout this chapter.

Web search engines discover pages by 'crawling' the Web, discovering new pages by following hyperlinks. Access to particular web pages may be restricted in various ways. For example pages on a corporate intranet may be visible only to those who can authenticate themselves as company employees. Web publishers may also impose additional restrictions on access by search engines, using the `robots.txt` Robots Exclusion Protocol (Koster, 1994). They may do this to prevent load on their servers, to help search engines avoid wasting resources crawling ephemeral or low value content, or for commercial reasons. Finally, search engines can, in general, only crawl pages which are linked to from pages already in the crawl. A page with no incoming links is extremely unlikely to be indexed by a general Web search engine.

The set of web pages which can not be included in search engine indexes is often called 'the hidden web', 'the deep web', or 'web dark matter' (Bailey et al., 2000).

2.2 Web structure

Web pages are connected by two major structures, the link graph and the URL hierarchy. We describe basic characteristics of the structure here and then the techniques that arise from the structure in later sections.

The link graph is formed when Web pages hyper-link to each other. The source page contains a reference to the URL of the link's target page. In many link graphs, there is a power-law degree distribution, which means that the probability of a link having degree i is proportional to $1/i^x$. One large-scale study (Broder et al., 2000) found $x > 2.1$. The power-law distribution means that most pages have very few incoming links, but a few have a very large number.

An outgoing hyperlink is represented by an 'anchor' in a web page. Usually, when the page is displayed in a web browser, the anchor is highlighted. The person may click on it in order to follow the link. If the anchor is textual we refer to it as the link's 'anchor text'. Anchor text often provides a useful short description of the link target, The anchor text of all the incoming links to a particular page can be aggregated for retrieval purposes into a useful surrogate for or adjunct to the actual document text. Anchors may be interpreted as labels on arcs in the general link graph, permitting the creation of topic-specific

sub-graphs.

Incoming anchor text can permit retrieval of useful pages (e.g. images, or pages in another language) whose textual content doesn't match the query. For example, the incoming anchortext for the page `http://www.robotstxt.org` includes many different descriptive phrases, such as 'robots.txt' (most common), 'http://www.robotstxt.org/' (very common), 'robots.txt system', 'RobotsTXT.org', 'Robots Exclusion Protocol', 'Robots Text','robots', 'here', 'robots.txt (Robots Exclusion Standard)', 'robots.txt file', 'robots.txt protocol', 'Martijn Koster's site about the Robot Exclusion Protocol.', 'Le protocole robots.txt', 'Crawler', 'den sedvänja som gäller på internet', ...

The URL hierarchy is a separate structure from the hyper-link graph. Taking as an example, the URL `http://blah.dog.com/dir1/dir2/file.html`, we can say that any pages that are both part of `.com` are loosely related. Pages that are part of `dog.com` are more strongly related and two pages from the same host `blah.dog.com` even more so. Then those on the same host that share some part of their directory path, such as `/dir1/` are further related. Site entry points, which are often the desired result of navigational queries, tend to be shallow in a host's URL hierarchy. Huberman & Adamic (1999) found that the number of pages on each host also follows a power law distribution. Using sampling techniques Lawrence & Giles (1999) estimated that the mean number of pages per server was 289.

The link graph and URL hierarchy are related. For example, Bharat et al. (2001) found that 76% of links in each of three separate crawls are within-host.

2.3 User behaviour

Much of the work that makes web information retrieval distinctive has arisen from understanding user needs. Users can research topics of interest, which is a traditional information retrieval activity. However, web pages link to each other, the user may wish to reach a page so they can browse further. Web pages are interactive, the user may wish to visit the page so they can shop, book travel or search a database.

The seminal contribution in describing web user needs was the taxonomy due to Broder (2002), first presented during the Web plenary session at TREC-2000 and later published in SIGIR Forum. Three types of information need were defined. **Navigational** where the immediate intent is to reach a particular page or site. **Informational** where the user seeks information relevant to their topic of interest. **Transactional** where the intent is to perform some web-mediated activity. In a followup study by Rose & Levinson (2004), roughly 60% the queries were classified as informational, 25% as resource/transactional and 15% navigational.

Study of navigational search proved to be the most immediately fruitful, because it can be evaluated using a single answer. Even before the taxonomy was presented, a number of studies were underway with navigational evaluation (Singhal & Kaszkiel, 2001; Craswell et al., 2001; Bharat & Mihaila, 2001). Since

then there have been further experiments in the context of the TREC Web Track (Hawking & Craswell, 2005). Ranking methods that were developed to answer navigational searches are described in the next section.

2.4 User interaction data

Another important way in which web search differs from traditional informational retrieval is in the truly massive amount of user interaction data collected by the major search engines. The most popular Web search engines are believed to log of the order of a billion interaction records each day. (A gigabyte of query text is accumulated approximately every 60 million queries.) These interaction logs include the queries people typed and the result links they clicked on. From them, search engines can assign general popularity ratings to pages (analogous to counting incoming links) (Culliss, 1999). They can also use clicks to associate queries with pages and then use the associated queries in ranking (analogously to anchor text) (Xue et al., 2004; Hawking et al., 2006). More detailed interaction data collected by the user's own computer can be used to improve search rankings (Agichtein et al., 2006).

Click patterns can be used to deduce relationships between pairs of queries and/or pairs of documents (Jones et al., 2006; Craswell & Szummer, 2007). Search engines may use click data as low cost relevance judgments for evaluating and tuning their systems (Joachims, 2002; Joachims et al., 2005). Interaction sequences can also be used to suggest spelling corrections or related queries.

Unfortunately, academic researchers have little access to search engine logs because of privacy concerns. Interaction sequences in logs may reveal a great deal of private information, *even if the data contains no usernames or IP addresses.* A good-faith attempt to make anonymised logs available to researchers in 2006, led to the unfortunate consequences described by Barbaro & Zeller (2006).

3 Three ranking problems

To be retrieved and presented to a user by a Web search engine, a page may have to pass three 'hurdles'. First the page has to rank sufficiently well in crawler prioritisation, otherwise it will never make it into the index. Second, if the system prunes its searches and uses a global ordering of pages, which is one technique for efficient query processing on very large indexes, the page must be high enough in the index to avoid being pruned. Finally, having been crawled and not pruned, the page must rank highly enough in the results list that the user sees it.

This gives three ranking problems. Effective ranking given a query has been studied for decades in the field of information retrieval (see Chapter 1), but in the web case specific techniques are needed to satisfy certain types of search. Here we focus on the navigational search type. The other two ranking problems,

in the crawl and in the index, are query-independent and have not been studied as rigorously. This section describes all three ranking problems.

3.1 Retrieval

Traditional relevance ranking technologies attempt to retrieve pages that contain relevant text. So for the query 'australian government' a good result is any page that talks about important aspects of the topic. However, another plausible scenario for the same query is that the user is performing a navigational search, where there is one correct answer `http://www.australia.gov.au`. In that case, different ranking techniques are required. Regardless of whether the user's need is informational, transactional or navigational, that site is likely to be a very good answer.

First consider the evidence available[1]. The goal is to get that page to the top of the ranking in a set of over 100 million matching pages (estimates are taken from Google). Using traditional ranking based on term frequencies (TF), assuming no stemming, the page has TF(australian)=2 and TF(government)=8. This is not strong enough to guarantee a top ranking for the page, as other pages may contain more (and a greater density of) occurrences of the query terms. New evidence is needed.

The two most successful sources of evidence for improving navigational ranking come from the link structure and URL hierarchy. The `www.australia.gov.au` page has almost 6000 incoming links which in the power-law distribution of the web means that this page is an unusually heavily linked. The URL signal that helps this page is the fact that it is a short URL and that it is the root page for a hostname.

Many incoming links for this page are likely to contain the words 'australian' and/or 'government' in their anchor text.

To be useful, query-independent evidence such as URL type, URL length, PageRank (Page et al., 1998), inlink count, click popularity etc. must be combined with query-dependent relevance scores derived from document text and with scores derived from external text such as URL words and anchortext. The best-known Web search engines are believed to use ranking functions which combine hundreds of features.

Chapter 1 presents a number of IR models which are capable of incorporating multiple sources of evidence. The reader is referred to Kraaij et al. (2002); Craswell et al. (2005); Ogilvie & Callan (2003) for web-oriented examples of combining evidence.

3.2 Selective crawling

Crawling can be thought of in terms of a queue, containing the URLs that the crawler has not yet requested. In the beginning, the queue contains a set of seed URLs. Then the crawler takes an URL from the queue, fetches that page

[1]These observations were made around May 2005, and may have changed since. However, they still illustrate the point.

(if available) and extracts its links. If the URLs have not been seen before, the crawler adds them to the queue. This can lead to the discovery of a massive set of pages, if unrestrained. Today large engines have crawls with billions of pages. In practise the crawl may not be strictly behaving as in queueing because of mechanisms used to achieve scalability, including distribution of the queue across multiple crawling machines and processes (Heydon & Najork, 1999), but it is still useful to think of it in terms of queueing. Recent large scale crawls for research purposes are described by Baeza-Yates et al. (2005) and Fetterly et al. (2003).

The scope of the crawl may be limited in a number of ways. The crawler may have a set of rules based on URLs which restrict the pages it covers, for example only crawling pages from a given list of hosts and domains. It may also have depth limits, only adding pages that are a certain number of link 'hops' from a seed page. Crawls usually also have an overall time limit. In theory it would be possible to continue the crawl until the queue becomes empty, getting 'all' the pages in a given scope. In practice, the time limit is useful because at some stage the crawl will become unproductive. The queue will contain only pages that are duplicates of existing pages, or perhaps pages from a very large site that is of low value. If there are crawler traps that generate an infinite number of interlinked URLs without useful new content, the crawl queue may never become empty.

Given that the crawl will stop early, with a non-empty queue, crawl priority becomes an interesting question. When the crawl terminates, a set of pages will have been crawled and a set of pages will have been missed. Which pages make it into the crawled set is determined by crawl priority.

The simplest crawl priority is given by a breadth-first traversal, which is equivalent to a first-in first-out crawl queue. This can also be thought of as a "generational" crawler, where the first generation is the seed set, the second generation is those pages of link distance 1 from the seed set and so on. Given that the seeds are important, it is not difficult to imagine that pages a shorter distance from the seeds will tend to be more important, under a basic locality assumption. Also, informal observations have shown that spam network pages and pages generated by an infinite crawler trap will be rare early in the crawl (assuming a good seed set) and will dominate later generations of the crawl. Najork & Wiener (2001) observed that crawling breadth-first also yields pages with high PageRank, which is further evidence that breadth-first traversal is desirable, in applications where crawling high-PageRank pages is desirable.

Another approach is to use a priority queue, rather than a fifo queue (Baeza-Yates et al., 2005). In that case, each URL in the queue is assigned a priority score, and the top-scoring page is dequeued. An important paper (Cho et al., 1998) studied such prioritisation in the context of the Stanford web, using in-degree, PageRank and match with a driving query. Chakrabarti et al. (1999) also considered crawling focused to a particular query as did a number of other studies (De Bra et al., 1994; Diligenti et al., 2000; Aggarwal et al., 2001). The general conclusion is that it is possible to effectively prioritise uncrawled pages, for example to selectively fetch pages that match a certain topic, based on the

content of linking pages that have already been crawled. This relies on anchor text that describes the target page and topic locality, that two pages on the same topic are more likely to be interlinked than two pages that are not (Davison, 2000).

The crawl queue prioritisation discussed so far was in terms of a newly created crawl. In the case of crawl maintenance it is possible to consider crawl priority and crawl updating as two separate concerns (Cho & Garcia-Molina, 2000). For crawl priority, mechanisms such as PageRank prioritisation or match with a driving query can still be used. The crawler must prioritise the crawled pages against the known uncrawled pages. Using numbers from Richardson et al. (2006), if we have a 5 billion page crawl, we can identify a further 20 billion URLs that are not yet crawled. So the selection problem can be seen as picking the best 5 billion from a set of 25 billion known URLs. The update problem is separate, involving revisiting the 5 billion known URLs to see if they have changed or been removed, and updating the index accordingly (Edwards et al., 2001).

In a general crawl, rather than a topic-focused one, PageRank seems to be the most promising method. Good navigational answers and high-PageRank are associated, as described in the previous subsection, so prioritising high-PageRank pages will tend to ensure that the navigational answers are selected first. If there are only one or two good navigational answers per site, and each site has on average 100 pages, then it's not surprising that PageRank-ordered crawling can get these top 1–2% of navigational answers. However, with a more sophisticated notion of which pages are desirable to crawl (many past studies have both prioritised using partial-crawl PageRank and evaluated using full-crawl PageRank (Cho et al., 1998; Baeza-Yates et al., 2005), it is possible that a much more sophisticated and effective crawl priority system could be developed. This system could take PageRank as an input, but also other aspects of the link structure, URL hierarchy and even page content.

3.3 Index organisation

There are many efficiency and engineering aspects of indexing and query processing which are critical to real Web search engines but which are not covered here. The reader is referred to Chapter 12 for a discussion of parallel IR and to Zobel & Moffat (2006) for a survey of state-of-the-art indexing techniques.

Recent papers (Long & Suel, 2003; Michel et al., 2005) have described a method for index organisation and pruning which works well for large-scale retrieval, but requires a global ordering of pages. The method is as follows. Rank all crawled pages according to some query-independent metric such as PageRank (Long & Suel, 2003), assigning document identifiers (docids) in increasing order. Then arrange each postings list in docid order. This is a global order because it is used to order all the postings lists, as opposed to other schemes such as impact ordering (Anh & Moffat, 2002). Use an AND query and skip lists (Pugh, 1990) so that for a multi-term query, the page-at-a-time processing can skip to pages containing all terms, avoiding processing of pages that only contain some

terms. Broder et al. (2003) show how the AND requirement can be weakened while maintaining efficiency.

Finally, given that PageRank is used in the results ranking as well as the index ordering, the pages that are processed first will be those with the highest potential score (due to their high PageRank component). This means that processing can be pruned (stopped) when it is decided that it is unlikely that any further pages will be found that outscore the pages already processed.

4 Other Web IR Issues

4.1 Stemming

Despite the consistent observation in TREC that moderately aggressive stemming, e.g. (Porter, 1980), tends to increase effectiveness, applying the TREC stemming approach in Web search is problematic for a number of reasons.

First, the chance of stemming errors is higher on the Web because the Web is multi-lingual and its vocabulary size is orders of magnitude larger than that of TREC, boosted by huge numbers of business names, product names, domain names, acronyms, placenames and names of people. It is feasible to recognize the language of Web documents with reasonable accuracy and thus to apply a language-specific stemmer during indexing. However, recognising the language of a one or two word query is far less reliable, particularly when spelling errors are common, and where searchers often lack the ability to produce accented letters. Similarly, although one could selectively avoid stemming based on the case of a word, searchers are notorious for failing to supply reliable case information in their queries.

Second, stemming errors are more likely to change the overall meaning of a query when queries are only one or two words. One of the present authors is particularly sensitive on this subject, his name having been confused by at least one search engine with a football club (the Hawks).

Third, stemming errors resulting in non-matching documents at the top of a search engine ranked list are more visible than in TREC ad hoc. In the latter, one may need to read the whole document to be sure it is irrelevant. Furthermore, a completely off-topic document returned a the top of the ranking has only a tiny depressing effect on average precision, particularly when averaged over 50 topics. In contrast, the same misfortune in a Web search is made obvious through display to the searcher of the result's title, URL and highlighted query words. Anecdotally, searchers view such failures with amusement, frustration, or bewilderment.

In practice, Web search engines typically apply a very light form of stemming in which the benefit strongly outweighs the risk of error.

4.2 Treatment of near-duplicate content

Established IR collections include instances of duplicate or overlapping content (such as different versions of the same news report) but on the Web the problem is magnified by the presence of systematic causes of duplication and by the resource implications of unnecessary crawling and indexing.

Certain websites set up to distribute information or software may be subject to excessive load at peak times. Their content may be replicated on webservers in other countries to distribute the load, or for purposes of reliability or preservation. In some cases this is handled imperceptibly behind-the-scenes, the content always published from a single web address. In other cases, the replicated information is published on a different site or *mirror*, meaning that identical documents appear on the web with different URLs. e.g. `mysoft.com/downloads/` may appear also as `mirror.xyz.ac.uk/mysoft/downloads/`. The reader is referred to Bharat & Broder (1999) for a discussion of techniques for detecting mirroring.

Aliasing of hostnames and of pathnames within a host is frequently practised. For example, the webhost `jupiter.xyz.ac.uk` may also be referencable as `www.xyz.ac.uk`, meaning that all its pages are accessible via two different URLs. Similarly, some (but not all) webservers treat URLs case-insensitively, meaning that e.g. `www.xyz.ac.uk/Sociology/Intro.HTM` can equivalently be accessed and linked to by millions of case variants. Often the content of pages fetched from different but equivalent URLs is exactly identical and the equivalence can be detected by simple checksum matching. However, when documents are dynamically generated, equivalent pages may be published with slight additions or alterations. For example, a page may include its own URL, the date at which it was accessed or a visitor counter. In these cases, more sophisticated techniques must be used to reliably detect the equivalence. See Broder (1997) for a discussion of *shingling*.

Dynamic generation allows a fixed unit of content to presented in a profusion of different forms, potentially with different colours and fonts, different images and with different sets of links. This may happen when an organisation operates a number of separate businesses each with their own branding. A more extreme example is provided by the online book retailers who generate websites to sell books on behalf of the bookselling giant Amazon, receiving a small referral commission. They all extract the basic content from the same Amazon database but compete for traffic by packaging and presenting it differently.

Dynamic URLs often include one set of parameters which identify the basic content, another set which control how it is presented, and a session-id parameter which may affect what is presented or which may merely enable tracking of an individual's browsing activity, e.g. `x.y.z/gen.asp?docid=123&stylesheet=look3&sid=11293129`. Search engines typically attempt to simplify URLs to avoid unnecessary fetching of the same basic content. They may strip off session-ids and presentation parameters. This can dramatically reduce wasted crawling effort but may sometimes result in missed content.

Correct treatment of near-duplicate content affects search engine operation

in important ways. Undetected duplication wastes valuable crawling time, increases network traffic costs, consumes extra disk space, extends indexing time and potentially clutters up results lists with repetitions. Not only that but duplicates add phantom nodes to the Web graph with adverse consequences to graph-based ranking measures such as PageRank.

4.3 Spelling suggestions

Web search engines receive a significant number of misspelled queries. Some provide a very helpful, "Did you mean X?" service. Due to the previously noted multi-lingual, neologism-prone characteristics of Web publishing, it is not at all feasible to make spelling suggestions by approximate searching within a normal dictionary word list. Nor is it useful to perform simple-minded approximate matching against the full vocabulary list of the Web as Web authors, like Web searchers, are highly prone to spelling errors. Very few misspellings would be detected by this method and suggestions made could easily be foreign words or spelling errors.

Details of commercial search engine spelling suggestion algorithms have not been published to our knowledge, but it is very likely that they are based on analysis of query logs. Cucerzan & Brill (2004) describe and evaluate methods for spelling suggestions based on logs.

4.4 SPAM rejection

To quote Broder once more, Web search engines operate in an adversarial information retrieval environment. Many people and organisations who make money via the Web have a strong financial incentive to try to ensure that their sites appear prominently in Web search results. A search engine optimisation industry has grown up to service this need, providing tools and services and even running conferences on how to improve website visibility within search engine rankings.

At the benign end of the optimisation spectrum, Web publishers may be encouraged to use good publishing practices – ensuring that titles, anchortext and content match the queries for which the page is a good answer and creating simple site and link structures to encourage linking and facilitate crawling. At the other end of the spectrum, spammers pursue their clients' interests regardless of harm to the interests of searchers, search engines and other publishers. They try to have their target pages highly ranked against queries which bear no relation to the subject of the page. They deliver different content to crawlers than to site visitors (*cloaking*) and they set up forests of interlinked artificial sites to increase PageRank and anchortext scores.

If search engines published the techniques they use to reject or down-weight spam pages, spammers would immediately be at work on better methods to defeat them. However, it seems clear that search engines use a combination of automatic spam classifiers and manual blacklists to cut down on unnecessary crawling and to keep unwanted pages out of search results. A spam score may

be assigned to a page based on the presence of unusual features in the text, on its participation in unusual link structures, and on its link distance from known black sites [trustrank]. This score could be used in setting crawl priorities and in assigning global document numbers (see Section 3.3.)

4.5 Adult content filtering - genre classification

Web search engines typically provide family-friendly filtering of search results in an attempt to prevent the display of sexually explicit material to minors. Much of this material can be detected on the basis of the presence of certain words and phrases, but a more sophisticated classifier might take into account words in anchortext, words in URLs, links from other sexually oriented sites, and even the relative prevalence of flesh tones in images.

Sexually explicit pages are one example of a webpage genre. It is possible to build classifiers for other genres, e.g. Glover (2001), (see also Chapter 3) and potentially to use them either in normal ranking or in personalised or contextualised result rankings.

4.6 Query-targeted advertisement generation

Web search engine companies rely on advertising revenue to fund their operations and to generate profits. In the early days of the Web, advertisements were indiscriminately published on search result pages for a small per-view charge to the advertiser, but conversion rates were low and revenues declined. The introduction of targeted advertising by Overture and later Google showed that when advertisements are related to the search query, conversion rates are much higher and advertisers are willing to bid substantial sums of money for query triggers related to their business. At the same time searchers are happier, because fewer advertisements are displayed and those which are displayed are more likely to be useful.

Search engines must operate a complex infrastructure to support the auctioning of query words and the accurate billing of advertisers. There is also a science in optimising which advertisements are displayed in order to maximise revenue to the search company. These issues were discussed by Ribeiro-Neto et al. (2005).

4.7 Snippet generation

Results presented by early Web search engines provided a hint as to the content of each page in the form of a fixed snippet from the head of the text. Such snippets can be generated by means of a simple lookup. More recent search engines now provide one or two short excerpts from the document, selected on the basis of localised relevance to the query. Tombros & Sanderson (1998) describe a basic method for query-biased summarisation and evaluate its effectiveness but do not address the issue of efficient generation of such summaries. Turpin

et al. (2007) have investigated the use of static compression schemes, caching and sentence re-ordering in improving the efficiency of snippets.

For the major search engines, snippet generation poses significant computational load. Since ten results are typically generated per query, the processing of a hundred million queries requires the generation of a billion query biased summaries.

4.8 Context and Web Information Retrieval

Even if a web search engine returns search results which maximise utility across all the users who submit that query, a minority may be poorly served. For example, those searching for information on a Tony Blair other than the former UK prime minister.

Contextualisation of search results, as described in the chapter on Context and Information Retrieval, can be applied by Web search engines to address this type of case. At a simple level, major search engines now routinely detect the country from which each search originates (based on IP address) and use this to:

- set the language of the interface,
- present advertising relevant in the local market,
- offer spelling suggestions appropriate to the language, and
- bias search results toward sites operating in that country.

For example, in response to the query 'bank', a major Web search engine currently[2] returns Bank of America, Migros Bank, Barclays Bank, or Commonwealth Bank in first rank, depending upon whether the search is conducted in the US, Swiss, UK or Australian contexts. In each case, the search interface assumes that the searcher speaks the dominant language of that country and is most interested in advertisements targeting that market.

This is called *localisation*.

Personalisation of search attempts to provide search results tailored for an individual rather than a nation.

A search engine can keep track of a user's interactions (e.g. the queries they issued and the results they clicked on) using cookies and search logs and can potentially use this data to disambiguate or expand queries based on inferred interests or preferences.

However, much of the research on Web search personalisation assumes that the personalisation context is deduced at the client side and is communicated to the search engine in the form of an expanded or transformed query. Personalisation at the client side means that all of a person's local electronic information (e.g. office documents, emails, and web pages downloaded) as well as a full

[2]May 2008

record of their Web interactions can be mined for personalisation context. Teevan et al Teevan et al. (2005) and Chirita et al Chirita et al. (2007) claim significant improvements in the quality of search results as a result of client-side personalisation.

4.8.1 Risks of personalisation and localisation

Personalisation and localisation can make things worse if incorrect assumptions are made about the searcher's skills, preferences or tasks. A holiday in Brazil doesn't necessarily make an English woman proficient in Portuguese! Furthermore, despite relaxing in Rio, she may still want to perform searches oriented to her job or to her life in the UK.

A good search engine will allow searchers to override automated personalisation and localisations.

Personalisation of search results has potential privacy implications, as illustrated in the following example. A woman uses her husband's laptop to search the Web. For some strange reason, her innocent but ambiguous query produces results which all relate to an unfamiliar sexual practice . . .!

5 Evaluation of Web Search Effectiveness

Progress of academic research in Web IR in the 1990s was impeded by the use of an established but inappropriate evaluation methodology. The task, judgments and measures used in the first TREC Web Track (Hawking et al., 1999) were inherited from TREC ad hoc. Inherent in this methodology are assumptions which can be seen to be violated by the majority of Web searches:

1. The purpose of all search is to find documents whose text contains information relevant to the topic of the query.

2. The searcher's goal is to find as many relevant documents as possible. (Implicit in the use of average precision as the key measure in TREC-8.)

3. It is appropriate to regard as relevant documents which contribute very little information and those which repeat information contained in other relevant documents.

4. All relevant documents are of equal value.

These assumptions are most inappropriate when search is navigational in intent but, even when this is not the case, Web searchers typically prefer site homepages and pages which provide access to a service. It seems safe to assume that the majority of people who type the query 'passport' to government or whole-of-Web search facilities, want to obtain or renew a passport, rather than write an essay about passports! Webpages describing the history of passports or incidentally mentioning the use or loss of passports are of almost no value in this context.

The typical Web searcher clicks on only one or two results. A study of search logs for the Australian Stock Exchange website showed that only one search in 35,000 resulted in clicks on all of the first ten results.

The evaluation procedures initially adopted in the TREC Web Track were unable to reveal any benefit for the link-based methods on which successful commercial engines rely.

Hawking & Craswell (2005) document the evolution of the Web Track toward greater Web realism through: engineering a more representative small corpus, making use of queries from search engine logs, and studying a series of deliberately Web oriented tasks.

5.1 TREC-9 Web Track: Realistic queries, rich link structure, traditional IR task

The TREC-9 "Main Web" task (run in 2000) went a long way toward Web realism in everything but the task. It used a corpus (WT10g) which had been specifically engineered to achieve a high degree of interlinking and to mimic other properties of the Web (Bailey et al., 2003). It also made use of queries selected from a public Web search engine log and augmented the normal TREC Ad Hoc relevance scale with a "highly relevant" category. The selected queries were slightly longer (3.4 words) than the reported Web average of 2.3 words.

Queries were selected by NIST assessors who imagined an information need behind the query and documented it in the description and narrative fields of the TREC topic statement.

For example, topic 488: 'newport beach california' was interpreted as

> **Description:** What forms of entertainment are available in Newport Beach, California?
>
> **Narrative:** Any document which refers to entertainment in Newport Beach is relevant. This would include spectator and participation sports, shows, theaters, tourist attractions, etc.

This interpretation is only one of the possible "needs behind the query". Other searchers posing this query might be interested in getting a general overview of the place where their boyfriend goes for family holidays, or in finding: the nearest airport to Newport Beach, photographs of Newport Beach, history of Newport Beach, accommodation in Newport Beach, driving instructions, a map, real estate investment potential, retirement homes etc.

In TREC-9, 35 documents were judged relevant to 'newport beach california' and one was judged highly relevant. The latter was `www.commpro.com/anaheim/tourism/beaches.html`. It contained information about Newport, Laguna, Huntington, San Clemente and Balboa beaches, but no Newport specific links. The only specific information about Newport Beach was:

> At Newport, there's also no shortage of sights and activities. The Upper Newport Bay Ecological Reserve is a 700-acre site that serves as home to wildlife and migratory birds. The protected coves of

> Corona del Mar Marine Life Refuge also offer tide pools alive with sea urchins, starfish and octopus. And at Corona del Mar State Beach, barbecue pits, showers, a snack bar and restrooms offer an ideal place for a family outing. And, early risers can watch as the fishermen of the Dory Fishing Fleet set out their catches of the day for sale at the foot of McFadden's Pier.

This would be a poor result for a search engine to return at rank one, because it fails to address the needs of searchers who are not seeking leisure activities. Even for those who are, there are no links from which accomodation or tours may be booked and none to maps or driving instructions.

In theory at least, successful Web search engines would attempt to maximize the utility of their search results page, summed over all searchers.

This might be achieved by including *portal* pages providing search-and-browse access to all or most aspects of the topic, such as site entry pages or hub pages oriented toward the broad topic. It would be appropriate to include results which are useful for more specific searcher needs but diversified across the range of likely needs. In these more specific cases, too, access to services, pictures and links to other resources may be more useful than paragraphs of informational text.

At the time of writing, the three largest Web search engines all return `www.city.newport-beach.ca.us` as the top ranked result. It is published by the Newport Beach city council and provides links to comprehensive and authoritative information about just about every aspect of Newport life and commerce, but no informational text. This page was not included in the WT10G collection but even if it had been, it is unclear how it would have rated with respect to the task as defined.

We suggest that the reader try this query themselves on their favourite whole-of-Web search engine, noting the characteristics of the results returned and judging them both, against the range of needs which might prompt a person to pose that query, and as if the task were to find paragraphs of relevant text.

5.2 Evaluation using web-specific tasks

In the TREC-2001 Web Track, the same corpus was used as in the previous year, but a task was introduced which makes no sense outside a Web context – given a name as a query, find the homepage of an entity of that name.

In this task, retrieval methods using anchortext and URL priors were shown to clearly outperform methods based on text content alone.

Subsequent web track tasks included named page finding, where web-specific methods were found to bring some but not as much benefit, and topic distillation, where the task was to find entry pages of authoritative sites relating to a broad topic.

5.3 Future directions for Web IR evaluations

The Web-oriented evaluations conducted in the TREC Web Track are closer to Web reality than the earlier attempts to apply TREC ad hoc methodology. However, they represent a substantial oversimplification.

A person may enter a query such as 'newport beach california' without knowing whether there is in fact a website devoted to that topic. It may be true that such a person obtains greatest value from home pages of official Newport sites but they may still gain substantial value from other types of resource, such as recent news reports and individual pages from which Newport holidays or accommodation may be booked.

In reflecting the value of a page of search results to an individual, a scoring system may need to encompass a very wide range of utility values for individual pages, and to sum scores only as far down the ranking as the person is prepared to look, while ignoring or even penalising results which are not useful to that person or which provide information or services which are similar to those already identified.

In evaluating the performance of a search engine on this query, as opposed to its value to an individual or to a class of similar individuals, it is necessary to sum its individual ratings over all the individuals who might pose the query.

5.4 Comparing results lists in context

Most measures employed in web and general IR, rate a results list by summing relevance values (possibly discounted (Järvelin & Kekäläinen, 2000)) previously assigned to individual pages in isolation. This approach has the strong advantage (at least in moderate sized collections) that judgments are re-usable. However, the score assigned to a result list in this way may substantially overestimate the actual value to a searcher:

- by assigning individual page scores which don't match the searcher's own assessment,
- by double-counting duplicate pages or pages with overlapping content,
- by counting pages which present information the searcher already possesses,
- by presenting information about aspects or interpretations of the topic which don't contribute to satisfaction of the searcher's goal, and/or
- by counting pages further down the list than the searcher is prepared to look.

Thomas & Hawking (2006) propose an alternative method in which pairs of results lists are compared side-by-side by searchers in real-life situations. In one experiment, volunteers substituted the side-by-side search interface for their normal Web search engine, whenever they conducted a search in the course of

their work, study or leisure activities[3]. After they entered a query, they were shown two sets of anonymised results, one from search engine A and one from engine B. Assignment of A and B to left and right was randomised. The searcher was given the option to record a preference for the list on the left or the one on the right or to indicate that the two lists were of equal value. A and B were then compared on the basis of the number of searchers who overall favoured one engine versus the other.

Although comparisons between pairs of results lists are not reusable and may not be sensitive enough to detect very small differences, this evaluation methodology has a number of advantages. Judgments are of naturally occurring information needs, conducted by the actual searcher, in the full context of the need, using whatever judging criteria matches their purpose in conducting the search.

5.5 Evaluation by commercial Web search companies

Web search engine companies have strong commercial incentives to evaluate the quality not only of ranking algorithms but of result presentation, summary generation, spelling suggestion and, particularly, ad. placement systems.

Tuning of ranking functions involving hundreds of variables, such as those used by the major Web search engine companies, requires sophisticated evaluation over very large sets of queries. It is well-known that these companies invest considerable resources in obtaining judgments for both competitive analysis and tuning.

Taylor et al. (2006) describe a method for optimizing a function of 375 variables using 5-level relevance judgments for training sets for 2048 queries, and speculate that even larger sets are used in commercial practice.

As mentioned earlier, search engine companies are also able to use the vast amounts of user interaction data (click logs) they accumulate as low-cost relevance judgments. Such judgments may be in the form of relevance scores for individual documents or as preference judgments between pairs of documents. As noted by Joachims et al. (2005), when a person scanning a list of search results skips over a document D1 and clicks on another (D2) at lower rank, this is a strong indication of preference for D2 over D1.

Search engine companies are able to randomly assign a fraction of the user population to a 'flight' which receives search results generated or presented in a non-standard way. Comparison of interaction data across flights can be used to estimate the value of a new feature or a new ranking algorithm.

6 Summary

In this chapter we have attempted to convey the key differences between IR on the web and IR in other settings, starting with the challenges of identifying and prioritising candidate documents for inclusion in the index. Next, we noted that

[3]Participants were, of course, able to opt out for any particular search.

web pages are interrelated, both by hierarchical site structure and in a hyperlink graph, and pointed out that measures derived from page interrelationships can be used as prior probabilities to improve search result quality. The effectiveness of web ranking can also be improved through exploitation of textual descriptions (such as anchortext) external to the documents themselves. On the Web, such descriptions are prevalent and reliably identifiable.

One of the main reasons why different sources of ranking evidence are highly beneficial on the Web, is that Web searchers behave differently to those modelled in non-Web test collections. It is common for people to use Web search engines for navigational and transactional purposes which are not supported in non-Web retrieval settings. Major behavioural differences are even observed when the search intent is informational: Searchers may value entry pages to authoritative sites on a topic and pages with useful topic links above documents containing chunks of relevant text.

Different user behaviour requires different evaluation methodologies, such as the ones mentioned in this chapter, and encourages the development of new ranking methods. It is now routine for Web search engines to *localise* search results to the country and language deduced from the searcher's IP address. Search results may also be personalised, either at the client side or using interaction histories maintained by a Web search engine.

We have briefly covered a number of topics, such as snippet generation, spelling suggestions, exploitation of user behaviour data, advertisement triggering and so on which are by no means restricted to web environments but which have been developed further or are currently more heavily used in the web context.

Further reading: In this brief chapter, it has only been possible to provide a cursory overview of a broad topic. Readers are encouraged to read the cited references in areas of particular interest and to seek out other relevant papers in recent proceedings of conferences such as World Wide Web (WWW), Web Search and Data Mining (WSDM), SIGIR and CIKM. In addition to well known Information Retrieval journals such as ACM Transactions on Information Systems (TOIS), Information Retrieval, Information Processing and Management, two recently established ACM journals, Transactions on the Web (TWEB) and Transactions on Internet Technologies (TOIT), publish a proportion of Web IR papers. Chapters 19–21 of a recent CUP textbook (Manning et al., 2008) provide more detail on some aspects of Web search, while Langville & Meyer (2006) covers the PageRank and related algorithms in considerable detail.

7 Exercises

1. Draw an architecture diagram of a breadth-first web crawler. Focus on data structures. How could a depth-first crawler work?

2. Try to characterize the behaviour of a major Web search engine by submitting queries of different types (e.g. a single stopword like 'the', queries

of different lengths, queries with spelling mistakes, queries for which there is an obvious right answer, ...) and observing the results:

(a) Does the search engine tailor results to the country it thinks you are in?

(b) Does the estimated number of results increase or decrease as you add words to a query?

(c) Can you guess the algorithm used to match advertisements to the queries you submit?

(d) Does the search engine ever eliminate stopwords from your query?

(e) Does the search engine appear to make use of stemming?

(f) Are the top ranked results for a single word query the documents with the greatest density of occurrences of that query word?

(g) How consistent are search results? If you submit the same query to the search engine in five minutes time, tomorrow or next week, do the answers vary?

(h) How frequently are websites crawled? (Try the query 'current time glasgow' and look at the result for a current time site like `www.timeanddate.com`.)

(i) How well does the spelling suggestion mechanism work when you type a misspelled query?

3. Contrast evaluation using judged queries vs evaluation using click logs. Where/how can mismatch between eval and real users arise?

References

C. C. Aggarwal, et al. (2001). 'On the Design of a Learning Crawler for Topical Resource Discovery'. *ACM Transactions on Information Systems* **19**(3):286–309.

E. Agichtein, et al. (2006). 'Improving web search ranking by incorporating user behavior information'. In *Proceedings of ACM SIGIR 2006*, pp. 19–26, New York, NY, USA. ACM Press.

V. N. Anh & A. Moffat (2002). 'Impact transformation: effective and efficient web retrieval.'. In *Proceedings of ACM SIGIR 2002*, pp. 3–10.

R. Baeza-Yates, et al. (2005). 'Crawling a country: better strategies than breadth-first for web page ordering'. In *Proceedings of WWW 2005*, pp. 864–872, New York, NY, USA. ACM Press.

P. Bailey, et al. (2000). 'Dark Matter on the Web'. In *WWW9 Poster Proceedings*.

P. Bailey, et al. (2003). 'Engineering a multi-purpose test collection for Web retrieval experiments'. *Information Processing and Management* **39**(6):853–871.

M. Barbaro & T. Zeller, Jr. (2006). 'A Face Is Exposed for AOL Searcher No. 4417749'. *The New York Times* `http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000`.

K. Bharat & A. Broder (1999). 'Mirror, mirror on the web: a study of host pairs with replicated content'. In *Proceedings of WWW8*, pp. 1579–1590.

K. Bharat, et al. (2001). 'Who Links to Whom: Mining Linkage between Web Sites'. In *Proceedings of IEEE ICDM-01*, pp. 51–58.

K. Bharat & G. A. Mihaila (2001). 'When experts agree: using non-affiliated experts to rank popular topics'. In *Proceedings of WWW 2001*, pp. 597–602, New York, NY, USA. ACM Press.

A. Broder (1997). 'On the Resemblance and Containment of Documents'. In *SEQUENCES '97: Proceedings of the Compression and Complexity of Sequences 1997*, p. 21, Washington, DC, USA. IEEE Computer Society.

A. Broder (2002). 'A taxonomy of web search'. *ACM SIGIR Forum* **36**(2):3–10.

A. Broder, et al. (2000). 'Graph structure in the web'. In *Proceedings of WWW9*, pp. 309–320, Amsterdam.

A. Z. Broder, et al. (2003). 'Efficient query evaluation using a two-level retrieval process'. In *Proceedings of CIKM 2003*, pp. 426–434, New York, NY, USA. ACM Press.

S. Chakrabarti, et al. (1999). 'Focused Crawling: A New Approach to Topic-specific Web Resource Discovery'. In *Proceedings of WWW8*, pp. 1623–1640.

P. A. Chirita, et al. (2007). 'Personalized query expansion for the web'. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 7–14, New York, NY, USA. ACM.

J. Cho & H. Garcia-Molina (2000). 'The Evolution of the Web and Implications for an Incremental Crawler'. In *Proceedings of VLDB 2000*, pp. 200–209.

J. Cho, et al. (1998). 'Efficient crawling through URL ordering'. In *Proceedings of WWW7*, pp. 161–172, Brisbane, Australia.

N. Craswell, et al. (2001). 'Effective site finding using link anchor information'. In *Proceedings of ACM SIGIR 2001*, pp. 250–257, New Orleans.

N. Craswell, et al. (2005). 'Relevance weighting for query independent evidence'. In *Proceedings of ACM SIGIR 2005*, pp. 416–423, Salvador, Brazil.

N. Craswell & M. Szummer (2007). 'Random walks on the click graph'. In *Proceedings of ACM SIGIR 2007*, pp. 239–246.

S. Cucerzan & E. Brill (2004). 'Spelling correction as an iterative process that exploits the collective knowledge of web users'. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 293–300, Barcelona, Spain. Association for Computational Linguistics.

G. Culliss (1999). 'User Popularity Ranked Search Engines'. Presentation at Infonortics Search Engines Meeting.

B. Davison (2000). 'Topical locality in the Web'. In *Proceedings of ACM SIGIR 2000*, pp. 272–279, Athens, Greece.

P. De Bra, et al. (1994). 'Information Retrieval in Distributed Hypertexts'. In *Proceedings of the 4th RIAO Conference*, pp. 481–491, New York.

M. Diligenti, et al. (2000). 'Focused Crawling Using Context Graphs'. In *Proceedings of the 26th VLDB Conference*, pp. 527–534, Cairo, Egypt. Morgan Kaufmann Publishers Inc.

J. Edwards, et al. (2001). 'An adaptive model for optimizing performance of an incremental web crawler'. In *Proceedings of WWW 2001*, pp. 106–113, New York, NY, USA. ACM Press.

D. Fetterly, et al. (2003). 'A large-scale study of the evolution of web pages'. In *Proceedings of WWW 2003*, pp. 669–678, New York, NY, USA. ACM Press.

E. J. Glover (2001). *Using Extra-Topical User Preferences to Improve Web-Based Metasearch*. Ph.D. thesis, University of Michigan. PhD Thesis: `http://ericglover.com/papers/glover_thesis.pdf`.

D. Hawking & N. Craswell (2005). 'Very Large Scale Retrieval and Web Search'. In E. Voorhees & D. Harman (eds.), *TREC: Experiment and Evaluation in Information Retrieval*, pp. 199–232. MIT Press.

D. Hawking, et al. (2006). 'Improving rankings in small-scale web search using click-implied descriptions'. *Australian Journal of Intelligent Information Processing Systems. ADCS 2006 special issue.* **9**(2):17–24.

D. Hawking, et al. (1999). 'Overview of TREC-8 Web track'. In *Proceedings of TREC-8*, pp. 131–150, Gaithersburg, Maryland USA.

A. Heydon & M. Najork (1999). 'Mercator: A Scalable, Extensible Web Crawler'. In *Proceedings of WWW2*, pp. 219–229.

B. A. Huberman & L. A. Adamic (1999). 'Growth Dynamics of the World-Wide Web'. *Nature* **401**:131.

K. Järvelin & J. Kekäläinen (2000). 'IR Methods for retrieving highly relevant documents'. In *Proceedings of SIGIR'00*, pp. 41–48, Athens, Greece.

T. Joachims (2002). 'Optimizing search engines using clickthrough data'. In *Proceedings of ACM KDD 2002*, pp. 133–142.

T. Joachims, et al. (2005). 'Accurately interpreting clickthrough data as implicit feedback'. In *Proceedings of ACM SIGIR 2005*, pp. 154–161.

R. Jones, et al. (2006). 'Generating Query Substitutions'. In *Proceedings of WWW 2006*, pp. 387–396, Edinburgh, Scotland. ACM Press.

M. Koster (1994). 'The Web Robots Pages'.

W. Kraaij, et al. (2002). 'The Importance of Prior Probabilities for Entry Page Search'. In *Proceedings of ACM SIGIR 2002*, pp. 27–34, Tampere, Finland.

A. N. Langville & C. D. Meyer (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

S. Lawrence & C. L. Giles (1999). 'Accessibility of Information on the Web'. *Nature* **400**:107–109.

X. Long & T. Suel (2003). 'Optimized Query Execution in Large Search Engines with Global Page Ordering.'. In *Proceedings of VLDB 2003*, pp. 129–140.

C. D. Manning, et al. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Full text on-line at `www-csli.stanford.edu/~hinrich/information-retrieval-book.html`.

S. Michel, et al. (2005). 'KLEE: a framework for distributed top-k query algorithms'. In *Proceedings of VLDB 2005*, pp. 637–648. VLDB Endowment.

M. Najork & J. L. Wiener (2001). 'Breadth-first crawling yields high-quality pages'. In *WWW 2001: Proceedings of the 10th international conference on World Wide Web*, pp. 114–118, New York, NY, USA. ACM Press.

P. Ogilvie & J. Callan (2003). 'Combining document representations for known-item search'. In *Proceedings of ACM SIGIR 2003*, pp. 143–150, New York, NY, USA. ACM Press.

L. Page, et al. (1998). 'The PageRank Citation Ranking: Bringing Order to the Web'. Tech. rep., Stanford, Santa Barbara, CA 93106. `dbpubs.stanford.edu:8090/pub/1999-66`.

M. Porter (1980). 'An algorithm for suffix stripping'. *Program* **14**(3):130–137. reprinted in Sparck-Jones and Willett.

W. Pugh (1990). 'Skip lists: a probabilistic alternative to balanced trees'. *Communications of the ACM* **33**(6):668–676.

B. Ribeiro-Neto, et al. (2005). 'Impedance coupling in content-targeted advertising'. In *Proceedings of ACM SIGIR 2005*, pp. 496–503, New York, NY, USA. ACM Press.

M. Richardson, et al. (2006). 'Beyond PageRank: Machine Learning for Static Ranking'. In *Proceedings of WWW 2006*, pp. 707–715, Edinburgh.

D. E. Rose & D. Levinson (2004). 'Understanding user goals in web search'. In *Proceedings of WWW 2004*, pp. 13–19, New York, NY, USA. ACM Press.

A. Singhal & M. Kaszkiel (2001). 'A case study in web search using TREC algorithms'. In *Proceedings of WWW10*, pp. 708–716, Hong Kong.

M. Taylor, et al. (2006). 'Optimisation methods for ranking functions with multiple parameters'. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 585–593, New York, NY, USA. ACM.

J. Teevan, et al. (2005). 'Personalizing search via automated analysis of interests and activities'. In *Proceedings of ACM SIGIR '05*, pp. 449–456, New York, NY, USA. ACM.

P. Thomas & D. Hawking (2006). 'Evaluation by Comparing Result Sets in Context'. In *Proceedings of CIKM 2006*, pp. 94–101.

A. Tombros & M. Sanderson (1998). 'Advantages of Query Biased Summaries in Information Retrieval'. In *Proceedings of ACM SIGIR 1998*, pp. 2–10, Melbourne, Australia.

A. Turpin, et al. (2007). 'Fast Generation of Result Snippets in Web Search'. In *Proceedings of ACM SIGIR 2007*, pp. 127–134.

G.-R. Xue, et al. (2004). 'Optimizing web search using web click-through data'. In *Proceedings of ACM CIKM 2004*, pp. 118–126.

J. Zobel & A. Moffat (2006). 'Inverted files for text search engines'. *ACM Computing Surveys* **38**(2):1–56.