Search engine has its root in information retrieval, which had undergone over half a century of research. How is information retrieval evolved into the search engines we see today?

1 Is it Information Retrieval? Is it Text Retrieval? No, it is Search

Information retrieval (IR) is not a household term. To most people not in the field, IR is simply too broad to be useful for understanding what it is. When you retrieve your last year's tax form from your steel cabinet or get a book from a shelf in the library, you are performing IR activities although they are done manually on some physical objects in the physical world.

To researchers and professionals working in the IR field, IR has a fairly specific meaning. By and large, it refers to IR activities related to unstructured textual documents such as books, reports, papers, documents, etc. Thus, IR is often referred to as *document retrieval* or *text retrieval*. For example, looking through the bibliographical records in a library is an IR activity, even though the search is only on some key fields or metadata, such as book titles, date of publication, etc. If the search is performed not just on some key fields but on the entire body of text, we call it *full-text retrieval*.

Interestingly, searching on the computer for employee records, product data, etc., is in general not considered IR even though search is also conducted on structured key fields. This is because the former leads you to books but the latter leads you to people. Therefore, the type of data you are searching for is important in characterizing if the searching activity is IR or not. The classification we discuss here is primarily used for restricting the scope of this writing and should not be treated as a general classification. For example, multimedia information retrieval, which considers both text and non-text elements in documents, has been actively studied in the IR community, but it is not covered here.

Today, decades of information retrieval research has matured into the science of *search engines* – a term that came into existence in the early 90's when the web became large enough to require the service of search engines to locate web pages on it. *Search engine* has become a much more popular term and carries a more precise meaning than "information retrieval" in the industry.

2 Applications of IR

2.1 Library systems

Information retrieval was one of the first applications of computers. Back in the 50's when computers were invented, computers were mostly used in two areas: scientific computation, most notably encryption and decryption, and the management of information in electronic form. Naturally, one of the places where a

large amount of information is managed is the library, where information about books, journals, etc., are catalogued, stored and retrieved. Thus, information retrieval was researched by library scientists more than by computer scientists (at that time computer science has not been established as a scientific discipline yet). In the 60's, as computers became more and more powerful, they became not only feasible but indispensable tools for automatic processing of large amount of information, which required the efficient storage and retrieval of information. Hence, information retrieval gradually became an important branch of computer science.

2.2 Search engines

Web search

Nowadays, most people get in touch with IR through web search engines. The usefulness of *web search* is beyond question. According to comScore, Google, Yahoo! and Bing performed a total of 16 billion *US explicit core searches* in May 2011, or about 6 million searches per second.[1] These commercial search engines index tens of billions of pages on the web[2] and make them searchable to everyone on the Internet.

Some of the major challenges of web search are:

- Web pages not only have different raw data formats but are written by different people for different purposes. For example, a nuclear scientist and a high school student may both write an article on the impact of nuclear weapons on humankind, but the former may be writing a report for the United Nations and the latter may be writing a term paper. Clearly, when a user queries for the impact of nuclear weapons on humankind, he/she may find one or the other interesting but not both. In other words, while search engines can pick up keywords, they cannot yet analyze the intent of the writers and target audience of the writings.

- Most people only type in a few keywords and expect to get useful results from the search engine. This causes a couple of problems. First, keywords could be ambiguous. Thus, it is difficult for the search engine to understand the meaning of the query. Second, and even worse, even when the query semantic is clear, it is hard for the search engine to decide the types of information that are most suitable for the users. Consider the previous example but from the user's end. How can we know if the user is interested in a formal, lengthy report or a casual writing on the topic? To resolve these problems requires deeper analysis and classification of information and continuous monitoring of the user's background interest.

[1] For detailed figures and the definition of "US explicit core search," see the comScore press release.

[2] `http://www.worldwidewebsize.com/`

Site search

While a web search engine searches the entire web for everybody, *site search* is dedicated to searching a corporate or intranet website. A good site search engine must take into consideration the site's structure, business and target audience. Although most web search engines can filter results based on URLs, they cannot capture the fine structure of a website and adapt the search results based on it. In addition, site search is required to meet certain unique challenges:

- Importance of a web page is determined by business factors, not by the keywords and links in the web page. For example, if the Chairman's message more important than a product announcement or white paper?
- While web search focuses on the precision of the search results (the returned results are relevant to the query), site search requires both high precision and high recall (all important results are returned to the users).
- Searching must be able to match the vocabulary used in the website to the vocabulary used by the user.
- Knowledge-based tools are required to help the users with refining their queries and exploring the website (e.g., providing the user a taxonomy of the terms used in the business).

Desktop search

Desktop search searches files on a desktop. Since files on desktops are organized in a folder hierarchy, desktop search must traverse the folder hierarchy and index the files in it. Desktop search must be able to index a large variety of data formats (Office files, videos, etc.), support search on a large number of metadata fields (e.g., last creation date, last update date, last read date, owner name, file name, file length, etc.), organize the search results based on fields and allow users to sort the results based on different criteria.

Media search

Media search refers to search engine that comes with the content on a storage medium. For example, a CDROM may contain 600 hundred Mbytes of text. The user cannot be expected to browse the content folder by folder. A search engine is needed, but how is it different from other types of search engines?

One notable aspect of media search is that the media is removable. In other words, the media can be plugged into and removed from a computer anytime. If the index is built on the desktop or on a server, when the media is plugged into a new machine it takes a lot of time to build the index; when the media is removed the index will become invalid. Thus, the index of the content must be built and stored on the media, meaning that the search engine must also be stored on the media. Thus, when the CDROM is inserted into a drive, the search engine is ready to be used. Media search engines must pay attention

to the low speed of the media and the desktop itself and compatibility across different hardware and operating systems.

Mobile search

Searching on mobile devices is increasing important due to the proliferation of mobile phones. Since mobile phones have small form factors, the user interfaces have to be simple. However, it is not enough to just shrink the interface of a web search engine to fit the small display because mobile search have some unique requirements:

- Mobile users are more interested in physical objects, such as bus stops, gas stations, friends, cinemas, and restaurants, etc., than long web pages. These are all spatial objects. As such, location-based and spatial queries must be supported in mobile search. Examples of these queries are: Find the gas stations that are nearest to my current position and driving direction (nearest-neighbor or NN queries) or tell me if any of my friends are in this area (window or range queries).

- The small displays of mobile phones dictate even higher precision compared to web search. Very often, if comprehensible summary is needed for the user to make sense of the result, no more than a couple of results can be displayed on a page and they better be the most useful ones for the user.

- In addition to the lack of easy input method, queries on mobile devices are often spontaneous. The user thinks of something and wants to find it. The query is expected to be even shorter than those on web search. Semantic matching is a must on mobile search.

- Even better, instead of requiring the user to input queries, mobile search should be proactive. In other words, the needs of a user are profiled and information is pushed to the mobile device and organized according to the user's interest profile.

2.3 Differences between library systems and search engines

Compared to Web search users, users of library systems are often professional librarians and as such their queries are more specific and less ambiguous. Thus, compared to search engines, library systems have some unique characteristics. We use the screenshot in Figure 2.3 to illustrate some unique features of library systems:

- Field search is essential in library systems for users to precisely control the results. While web search engines support field search (page titles and metatags, field search is not essential and is rarely used by the users.

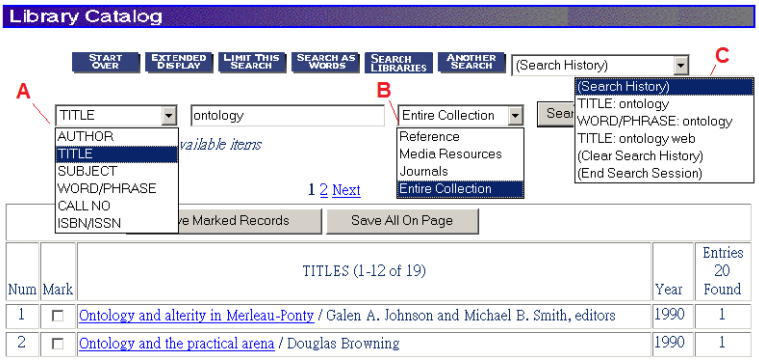

Figure 1: A library system's search interface: field search (A), collections (B), and search history (C).

- Collection boundaries are clearly defined in library systems, allowing a query to be applied to one or more specific collections (e.g., publications from a particular publishers, publications in a particular subject, etc.). Search engines can restrict search based on URLs, but URLs cannot clearly specify a collection's boundary.
- Search history is an *essential feature* of library systems. Users not only can view past queries but can operate on the results of previous queries, e.g., perform an intersection on the results of two previous queries or subtract the result of one query from the result of another query.
- Since most library systems support the Z39.50 standards for the exchange of queries and results across different library systems, *federated search* can be easily supported.
- On the other hand, relevance ranking is not as difficult in library systems compared to search engines since publications can be precisely characterized by the titles and summaries, which are short and less ambiguous compared to web pages. Besides relevance ranking, ranking by publication dates and author names meet the requirements of most library users.

3 Questions

1. Compare the differences between a product search system and web search from three perspectives: the types of users using the systems, the types of data searched by the systems, and the functional requirements of the systems.

