Поиск информации: проблемы культурного наследия. Исторический взгляд

Cultural Heritage Challenges to Information Retrieval – A Historical Perspective

Пошук інформації: проблеми культурної спадщини. Історичний погляд

Матс Линдквист
Национальная библиотека Швеции, Стокгольм, Швеция

Mats G. Lindquist
National Library of Sweden, Stockholm, Sweden

Матс Ліндквіст
Національна бібліотека Швеції, Стокгольм, Швеція

Поиск информации – одна из сфер, где раньше всего начали применять то, что тогда называлось «автоматизированная обработка данных». Технологии и методы поиска информации совершенствовались вместе с развитием компьютеров, чтобы соответствовать требованиям доступности новых видов информации и необходимости их поиска.
Поиск хорошо структурированной и закодированной информации развивался и перешел к содержательному поиску по библиографическим записям, а затем – к более сложным видам документов, появившихся благодаря автоматизации офиса. С появлением Интернета как платформы и носителя пертинентной информации поиск данных столкнулся с новыми проблемами.
Материалы, относящиеся к культурному наследию, – это разнородный набор структурированных описаний и разнообразных объектов. Соответственно, возникают новые проблемы при поиске информации.

Information Retrieval (IR) is one of the earliest application areas for, what was then called, Automated Data Processing. IR technologies and methods have advanced in line with hardware developments and to meet challenges posed by the availability of and need to search new types of information.
From searching in very structured and coded information IR developed to content searching in bibliographic references, and then to the more difficult documents that came from Office Automation. When the Web emerged as a platform and carrier of pertinent information IR faced new challenges.
Cultural Heritage material is a heterogeneous set of structured descriptions and objects of a great variety. For Information Retrieval new challenges evolve.

Пошук інформації – одна зі сфер, де раніше за інші почали застосовувати те, що колись називалось «автоматизована обробка даних». Технології та методи пошуку інформації вдосконалювались разом із розвитком комп'ютерів, для того, щоб відповідати вимогам доступності нових видів інформації та необхідності їх пошуку.

Information Retrieval as a special branch of computing

Background

Earlier names for this application area reveal what the original context was: Literature search, or document retrieval, i.e. retrieving relevant bibliographic records from some kind of database. The nature of the material was mostly scientific and technical. Later bibliographic databases were built with richer records which included abstracts or summaries. These records were referred to as document surrogates.

Early applications started in the 1960's and have since developed to take advantage of advances in computing and data processing, what we today call Informtion Technology.

In the 1970's there was a growth of information retrieval services offering online searching on emerging telecommunication networks. External bibliographic databases and internal corporate document databases stimulated developments in IR. The late 1980's can be called «the late classic period» for IR which lasted into the 1990's.

The emergence of the WorldWideWeb in the mid-1990's changed our ways of creating, distributing and using information; for IR a whole new field of opportunities – and problems – opened.

Brief overview of IR developments (to the mid-1990's)

There are different aspects of the development of IR: Technological developments for data processing (i.e. Information Technology) made it possible to create and store greater volumes of information, and in a greater variety of forms. The need for effective IR increased, which stimulated developments in IR functionality. This, in turn, created a demand for further technological developments

The technology aspect:

The early applications were limited by both processing and storage capacity. Magnetic tape replaced punch cards, but for quite some time the record structure was «card images» and subsequently more generous fixed-field records. Magnetic tape was created to be used in computerized typesetting of (printed) indicies to scientific literature. Magnetic tape was also used for the distribution of catalogue records for libraries (MARC records). IR technology was developed to match queries and database records with a minimum of tape spins.

When disc storage gradually became available it gave the possibility of direct access and IR techniques were revolutionized. However storage was very expensive and online searching was an exclusive activity.

The continous development is storage capacity and decreasing cost for storage led to a growth in online search serivces and internal business applications.

The Content aspect:

The early applications for IR were based on highly coded and structured records. The mental model was list entries (lines of codes plus text data), which then emerged into fixed-field bibliographic references – logical catalogue cards.

Computerized typesetting of [printed] indexes, or abstract journals made the bibliographic records available in machine-readable form that could easily be distributed on magnetic tape. This led to a growth of information search services in the 1970's.

With magnetic tape as the information carrier the physical limitations were lifted: variable length fields and records began to emerge. The records could be longer and contain longer texts such as abstracts. There were also fewer restrictions on the content, e.g. regarding the number of authors and the length of names and other literals.

Various collections of computerized texts started to emerge, usually within the context of linguistic research. And then, when word processing became integrated with data processing, the volume of machine readable documents started to grow exponentially.

Office Automation added word processing to the scene of data processing. Large amounts of relatively large texts, office documents, became available and a target for IR. Control of the structure and preparation of the material for searching made it possible to achieve both high precision and high recall.

The late 1980's can be called «the late classic period» for IR which lasted into the 1990's.

There were plenty of developments in functionality, efficency in searching, and flexibility in results reporting and post-processing. Office documents became important electronic resources and Document Management systems were developed; IR, sometimes called document retrieval, became an integral and fundamental part of them.

The functionality aspect:

IR is about matching a search request (query) with records (documents) in a database and present a result set of records (hits). The classical performance measures for IR are precision and recall.[1] How the database is built gives the prerequisites for searching. The most important activities are selection of material (what to include in the database) and checking or creating the structure of the records (how to partition the information and assign field labels to the parts). To choose which database to search is indeed often the first iteration of IR.

Search queries are expressed in a search language consisting of terms and relationships between terms. In an historical article Hahn [1998] divides search capabilities into two categories. The first contain

[1] Precision is the percentage of relevant hits of the total number of hits. Recall is the percentage of relevant hits of the total numer of relevant documents in the database.

those that help to specify the relationships between terms in a query; here we find the Boolean operators (AND, OR, NOT) and proximity operators (NEAR, WITHIN SENTENCE, WITHIN n WORDS, PHRASE, etc). The second category contain those that facilitate the interpretation of a particular word; here we find truncation and wild-cards (forms of masking individual characters in a word at the end, in the beginning or in the middle), fuzzy search, numeric and date ranging. In this category we also find fielded search (directing the query to a specified field of the records, such as title), term weighting, and synonym expansion. Fielded search is perhaps the strongest capability to improve precision.

Then there are other search capabilities that relate to the search method. Here we find iterative search –»the capability to further modify the results from a previous search» (requires a result sets history to be available). This feature makes precision better. Then there is functionality to improve recall: Fuzzy searching, vocabulary browse and relevance feeed-back.

So at the end of «the late classic period for IR» there were systems with very sophisticated search capabilities designed for searching in databases of large and diversified documents. Users were empowered to control the search process through iterative search functionality. IR capabilities were an important part of Document Management systems.

The challenges addressed were primarily the enhancement of precision when searching in longer texts, and increased user control of the search process.

Emergence of the web and new IR directions

The internet emerged in the mid 1990's. In some cases it was just the access route to databases, but the web quickly grew to become both the communication channel and the application platform for many uses. So the web was not only a dissemination channel, but also effectively took the role of information carrier and storage system (physically disc-based).

Information on the web, ususally referred to as resources, has no common structure. Some are «document-like» but others can only be described imprecisely as «web pages». They may contain very different kinds of material in addition to text: images, sound, video.

The challenge for IR was how to deal with no structure in a high volume landscape. There is generally a lack of metadata so fielded search is not developed. There are some «fields» like language, sometimes date, material type (e.g. images, video), URL:s, but no metadata that relate to the subject matter or content. And there is no central authority to provide metadata control so web resources will continue to suffer from the lack of metadata.

Outside of the web a new material type started to emerge at this time – the fulltext scientific article packaged in online journals. They were offered with very restricted access and they were accessed and searched in the same way as traditional bibliographic databases.

The searching techniques on the web seemed to start from a position of no experience: quite often very primitive user interfaces met the web searchers. It was as if 20 years of developments in IR functionality was forgotten (or ignored).

The traditional IR performance measures precision and recall are based on the notion of relevance (of retrieved documents); in the web environment these measures are difficult to establish. With the very large result sets that are typical of web searching relevance ranking is of the utmost importance. But to base it on the classical precision measure is difficult. So the search engine makers are redefining relevance [Brooks, 2004]. Most attention has been given Google's ranking algorithm. Ranking has become one of the fundamental capabilities of web search engines. But the ranking algorithms are not transparent, and the user has no control over the tuning of the results.

«Lately, however, there have been signs that our honeymoon with purely ranked retrieval system is coming to a close. Bing, Microsoft's recent re-launch of its web search offering, touts itself as a 'decision engine' rather than simply an engine to match search queries to documents. More substantially, Bing's interface offers users a variety of ways to interact with the search engine. --- Yahoo!'s Search Assist provides real-time query suggestions to users as they enter search queries. On the web, we are seeing the initial signs of search engines engaging users in a more interactive query elaboration process.» [Tunkelang, 2009]

In terms of classical performance measures the challenge for web searching is to improve precision – e.g. by metadata enhancements (without sacrificing recall). There is, however, no natural actor to provide control over the metadata models, and the scope of the web makes any attempts bound to failure.

The search method that has evolved is free text retrieval combined with faceted search. This is especially obvious in Enterprise Search applications where often the facets are subject to manual metadata enhancements.

Summary of the consequences on IR from the emergence of the web:

- search functionality was primitive at first, developed somewhat during the years, but still lack the strength of fielded search (metadata based) and iterative search (based on result sets) among other search functions.
- the concept of relevance was difficult to operationalize and it was replaced by non-personal ranking of results.
- intranets developed inside corporations and IR became part of Enterprise Search systems.

Cultural Heritage applications – new IR challenges

Cultural Heritage applications, bringing together the offerings from libraries, archives, museums and audio-visual, are based on a fairly recent ambition. The different domains have traditionally not shared data or service offerings. The general trend of convergence of information, technology and applications has made cultural heritage portals (and other services) a natural development.

The web is now the way to make cultural heritage information available, as opposed to closed databases that are typical for institutional or commercial search services and internal business applications. The openness of the web is a prerequisite

With Cultural Heritage material we mean databases from cultural heritage institutions such as museums, libraries, archives and audiovisual archives. These databases consist of structured data, but there are many different structures.

«In the Cultural Heritage domain, where much of the information is from multiple (and potentially conflicting), distributed, biased and historic sources, imperfect information is pervasive. The nature of such imperfections can take a number of forms, including:» [Clough et al., 2009; the list entries below are abbreviated]:

- Missing (unavailable or incomplete)
- Uncertain (the accuracy is not known)
- Imprecise (is of varying degree of precision)
- Ambiguous: where the correctness of a particular value may depend on the context)

There is also the possibility that there is a scientific debate about what the «truth» is.

The challenge for IR is to manage interoperability and to improve recall (since imperfect information has a negative effect on recall).

Normally fielded search is a means to improve precision. However, in this situation, with diverse structures, fielded search will improve recall if the differences in structure can be overcome. A straightforward way to homogenize structures is to design a mapping model. This means the creation of a «superset» of metadata and rules for how to map into this set.

An example of this is Europeana[2] where contrubuting institutions map their metadata to ESE – the Europeana Semantic Elements set, which is based on Dublin Core.

Then there is a more ambitious way: adopting (or participating in) the semantic web. Semantic information portals «are based on semantic web standards and machine «understandable» content, i.e. metadata, ontologies, and rules, in order to improve structure, extensibility, customization, usability, and sustainability of traditional portal designs.» [Hyvönen, 2007].

As opposed to the word based search of traditional IR semantic search aims at finding the concepts related to the documents at the metadata and ontology levels. Traditional IR capabilities such as free text retrieval and vocabulary search can be used as complementary techniques when bridging the gap between queries expressed in free form and the ontological concepts of the semantic structure. Free text searching can also be applied to the content of metadata elements, such as names and other short strings, and to summaries and descriptions of objects. This is similar to traditional IR in databases of document surrogates.

[2] The Europeana initiative by the European Union consists of many projects: Europeana version 1.0 and EuropeanaConnect are the main technology oriented projects. In addition there is a whole group of projects whose main task is to contribute content to Europeana, see http://version1.europeana.eu/web/europeana-project/

In the Europeana version 1.0 project a data model for semantic searching is developed.

The Europeana Data Model, EDM, is a qualitative change in the way Europeana deals with the metadata gathered from data providers and aggregators. It provides extra expressivity and flexibility compared to the Europeana Semantic Elements (ESE). In EDM a distinction is made between the intellectual and technical creation that is submitted by a provider (a bundle of resources about an object), the **object** this structure is about, and the **digital representations** of this object, which can be accessed over the web. EDM follows modelling principles of the Semantic Web. So there is not one single fixed schema that dictates just one way to represent data.

There is an experimental system in the Europeana portal. It is a research prototype of the semantic search engine (http://eculture.cs.vu.nl/europeana/session/search) and it gives some of the flavour of semantic searching.

Clough et al. [2009] have done experiments with semantic web technologies on the collection at the Tate Gallery in London. Their conclusion is: «It is clear that technologies from the Semantic Web have the potential to improve information access to cultural heritage collections. As the availability of publicly-accessible data in an interoperable form (e.g. as linked data) increases, the potential for linking and sharing cultural heritage material with other resources also increases.»

A summary of the challenges for IR:

For the classic period: search efficiency (primarily speed), and search functionality, primarily to improve precision when searching in longer texts; and increased user control of the search process.

For the web period (continuing): to manage absence of structure and large volumes, and to improve precision without loosing recall.

For Cultural Heritage applications: to manage interoperability and to improve precision and recall, and to embrace technologies from the Semantic Web and Open Linked Data.

Cultural material as web resources

There is, of course, much material that are a part of the cultural heritage freely available on the web but not in the databases from cultural institutions. To search in these resources new search tools will have to be developed in addition to some of the traditional. The Internet Archive, and the International Internet Preservation Consortium (IIPC) are working on these developments. In the future we might see special Cultural Heritage Search Services based on these archives in addition to the portals of the heritage institutions such as Europeana.

References

Clough, Paul, Neil Ireson and Jennifer Marlow, «Extending Domain-Specific Resources to Enable Semantic Access to Cultutal Heritage Data, J.of Digital Information, vil. 10, no. 6 (2009). Available at: http://journals.tdl.org/jodi/article/view/698/578 [2010-04-15]

Hahn, Trudi Bellardo, Text Retrieval Online: Historical Perspective on Web Search Engines, Bull. ASIST, April/May 1998, pp.7-10. Scanned version: http://www3.interscience.wiley.com/cgi-bin/fulltext/109862839/PDFSTART

Hyvönen, Eero, «Semantic portals for cultural heritage». Semantic Computing Research Group, TKK, 2007. Available at: http://www.seco.tkk.fi/publications/2007/hyvonen-portals-2007.pdf [2010-04-20]

Lindquist, Mats G., «Information Retrieval from KWIC to Enterprise Search», Presentation at the Online Information Meeting, London, 6 December 2007. Available at: http://www.slideshare.net/mglindquist/iolim-2007-12-06-m-lindquist [2010-04-06]

Terrence A. Brooks, «The nature of meaning in the age of Google», Information research, Vol. 9 No. 3, April 2004. Available at: http://informationr.net/ir/9-3/paper180.html [2010-04-15]

Tunkelang, Daniel, «Reconsidering Relevance and Embracing Interaction» Bull. ASIST, October/November 2009, Available at: http://www.asis.org/Bulletin/Oct-09/OctNov09_Tunkelang.html [2010-04-15]

Veal, D.C., «Techniques of document management : A review of text retrieval and related technologies», Journal of Documentation, vol. 57, no. 2, March 2001, pp. 192-217.

Links

Europeana – a research prototype of Europeana's semantic search engine:
http://eculture.cs.vu.nl/europeana/session/search

The Europeana Data Model is available at:
http://version1.europeana.eu/web/europeana-project/technicaldocuments/

The International Internet Preservation Consortium
http://www.netpreserve.org/about/index.php

The Internet Archive
http://www.archive.org/

The semantic web
http://semanticweb.org/wiki/Main_Page