Relevance ranking of results from MARC-based catalogues: from guidelines to implementation exploiting structured metadata

By Alison Dellit & Tony Boston

Abstract

The development of new search technologies, and new search behaviours, has opened up new opportunities for libraries to expose their holdings. Our users are more used to searching large aggregations of items, and to being able to obtain rare and unusual items relatively easily. Library holdings, if exposed through centrally searchable union catalogues like Libraries Australia, could see a significant increase in usage in the new "long tail" environment. To take advantage of these opportunities, however, libraries need to provide tools to make resource discovery easier across large stores of records - including relevance ranking and clustering of search results. During 2005, the National Library has taken steps forward in harnessing these technologies for Libraries Australia. An explanation of the development of a set of rules for the relevance ranking of bibliographic records is given, and the work done to advance clustering of relevance ranking and other potential search technologies is explored.

Introduction

In the last five years, the search behaviours of most Australians have undergone a profound change. The rapid uptake of Internet access - more than 50% of Australian homes now have internet access - has meant that more people than ever before can access and search vast stores of information from their homes or their workplaces [1]. This trend has both encouraged and been helped by the development of new search technologies; in particular, the excellent relevance ranking of Google has enabled users to find highly relevant web sites using a simple keyword search, without any obvious fielded searching. The development and popularisation of automated recommendation services, particularly by services such as Amazon, has enabled users to browse material in many different ways, not simply by following subject or classification trails. These changes offer new opportunities for libraries to expose our collections. If we can successfully harness the potential of this new environment, we can make resource discovery easier for our patrons and offer them the ability to identify items which may not be held locally.

Libraries Australia and the Long Tail

In an article in 2005 for *Wired* magazine, Chris Anderson argued that the Internet was challenging the idea that most revenue comes from a small number of best-sellers in most industries. He argued that, with online retailing and distribution, we are entering "an age of abundance" of choice: that online businesses are able to offer an enormous catalogue of material, and that people are responding by exploring diverse material. Hence, Anderson claimed, a substantial portion of profits for companies like Amazon and NetFlicks came from the "long tail" of their catalogues: the 80% of material that is accessed only rarely [2].

Since that article's publication it has spawned dozens of articles, and a few books, examining the concept and the implications for future business models for online retailers. Anderson has updated the figures in the initial article, has produced a quickly selling book, and runs a website, dedicated to discussing the phenomenon. Anderson's latest figures claim that 25% of Amazon's total sales come from rare or out of print items that are not available in off-line book chain stores.

Lorcan Dempsey has since argued that most libraries had failed to absorb the relevance of Anderson's argument to the sector, including the importance of libraries' aggregating the supply of material, and taking advantage of the aggregation of demand [3]. In Australia, Libraries Australia offers significant opportunities for us to do both. The launch in December 2004 of the new Libraries Australia search service with its Google-like search box and the launch in February 2006 of a free version of the service were steps towards this goal, as now all Australians, wherever they are, can search the holdings of Australian library collections through the Australian National Bibliographic Database.

Libraries Australia contains approximately 16 million bibliographic records, some of which represent rare, unique or obscure material. As Australian library collections are more widely exposed, a world of information is opened up: from indigenous histories based in Charles Darwin University Library, through to the vast music collections at the NLA.

Our challenge – as a library community – is to make these resources as easy to find and get as the best "long tail" businesses resources are. Finding and getting a library item should be no more complicated than searching and ordering on Amazon, or Ebay.

To do this, we need to make searching Libraries Australia as easy and intuitive as possible – including providing new ways for users to browse material; and we need to make getting resources as easy as possible. This paper reports on efforts to improve the searchability of Libraries Australia. Discussions on improving the getting of Libraries Australia material is outside the scope of this paper, however, we would like to note the recent establishment of the Rethinking Resource Sharing Reference Group, which is looking at this problem.

New search technologies: opening our collections

In a 2003 report prepared for the Library of Congress, Marcia Bates presented a detailed review of studies covering use of card and online catalogues. Using data from both card catalogues and online catalogues, she concluded that: "The average user identifies their search terms with the whole subject query. It does not occur to them that it might be called other things by the catalog. They look up their topic, do not find it, therefore the library must not have anything on it" [4 p.7]. She estimated that just a third of first time subject searches matched the assigned Library of Congress subject heading. She also quoted the blunt assessment of CR Hildreth, from a 1997 study: "users of this online catalogue search more often by keyword than any other type of search, and their keyword searches fail more often than not"[4 p. 9].

This trend will only increase as users bring expectations developed through Internet use into library environments. They will not click a drop-down box to work out which field they should be searching upon, nor do most people think about constructing a search

query using a machine-friendly syntax. Because large numbers of people expect results to be listed in order of relevance, they will assume that if there is nothing relevant on the first page, there are no relevant results. They will not try to get a result set small enough to browse, they will want the system to give them what they want, quickly and painlessly. Using new search technologies, we can meet this expectation.

Since it was first popularised by Google, relevance ranking has transformed the way that people search. Before relevance ranking was popularised, most search technology focused not on bringing relevant material to the top of the list, but on eliminating irrelevant material from the result set. This approach did not always make it easy to find material if the result set was large. It made it harder to search very large databases, within which many items might be somewhat relevant.

With relevance ranking, users do not have to despair of a large result set. In fact, most users do not notice how big the result set is. Instead, they notice whether the results listed in the top half of the first page provide what they want. Rather than scanning the whole list, users will jump straight to a "good enough" relevant result in the first few, or will choose one of the first few results as a starting point to browse for more relevant items. Clustering of results, by subject, author, genre format or date, can allow users to easily refine a search, with one click of the mouse. New methods of browsing – via peer recommendations, or through subject taxonomies, or related websites – have emerged. Users now browse along individually tailored paths, rather than a linear shelf order, and often browse a vast variety of material formats in one session.

During 2006, Libraries Australia implemented relevance ranking as the default sort order for both the simple and the advanced search as well as via Z39.50. We are currently working on a prototype that will include the clustering of search results, suggestions to users for better searches, and recommendations from users. These developments were the result of a project run in mid-2006 examining the best relevance ranking methods for bibliographic records.

Implementing Relevance Ranking in Libraries Australia

Standard relevance-ranking algorithms have been mostly developed for full-text documents. They generally emphasise an inverse frequency algorithm - the higher the proportion of the field or whole document that a search term occupies, the more relevant that match is. However, bibliographic records are not full text documents, and often relying on this algorithm is misleading. Bibliographic records often have short fields, such as additional authors or alternate titles that may repeat terms several times but not indicate a higher level of relevance than terms in longer fields. For example, a report by the federal immigration department might have several short heading containing the words immigration and ethnic, including entries for corporate name, subject, title and additional title, even though it is unlikely to be more relevant than other items whose records contain less fields.

If read properly, the highly structured nature of a bibliographic record provides a wealth of data to assist in relevance ranking, suggesting a different approach. There are several sources of data that we can use to decide which record is more relevant: there is the information in the bibliographic record itself, especially the subject, author and title fields; there is the holdings information, describing which libraries, and how many, carry an item; there is data derivable from a relationship between a searcher and the title, such

as reading age, there is availability information such as local holdings accessible by the searcher, availability and cost from bookstores, there is citation information and there is user-generated data, such as circulation data, user reviews, tags or recommendations. We were not able to incorporate and use all this data during the development process, and some we plan to explore in the future.

In the course of developing a set of rules to govern which items should be ranked higher in search results, we developed the following principles for the relevance ranking, based on the data in the bibliographic record and the holdings information.

1. Matches in the author, subject and title fields, and those fields which describe the format, nature or form of the item, are more important than general matches within the record.
2. Matches in multiples of the above fields are more important than matches in just one of those fields.
3. Where there is one or more query terms, an exact match of the term (where what was typed in to the query box is exactly the same as what is in a field in the record) is much more important than a phrase match (where what was typed in to the query box matches exactly a part of what is in a field in the record), which is more important than a word match (where all the terms in a query box appear in the field, but not necessarily next to each other).
4. The strength of the above point applies differently to matches in different fields. An exact match in the title field is much more important than a phrase match, which is much more important than a word match. In the author field, an exact match is slightly better than a phrase match, but both are much better than a word match. In the subject field, there is a slighter difference in importance between the three types of matches.
5. Matches in the main fields (in MARC, for example the 245$a,b; and the 100 field, for example) are more important than matches in additional entry fields, or alternate titles.
6. Matches of query terms as typed into the query box are more important than matches of a stemmed or spelling-corrected version of the query term.
7. All other things being equal, collection-level records are likely to be more relevant than non-collection level items.
8. Items with more holdings are likely to be more relevant than items with less holdings.

We used these principles as the basis for testing and configuring the Libraries Australia search engine. Relevance is a hard concept to test, because it varies between individuals and the tasks that individuals are attempting to perform. Given the early stage of the project, it was decided that usability testing was not the aim of the process. Instead, we used the considerable expert resources at the National Library and Libraries Australia – a group of reference librarians – to make initial judgments about relevancy. This team had a period of experimentation to work out a starting set of basic rules, and then ran three iterative tests on a standard set of search results that was culled from the most popular searches on Libraries Australia in May 2006. Between the test sessions the set of rules for each system was altered based on the results of the first test, and extra testing was conducted to refine the new set of rules.

In the first two rounds of testing, the team identified result sets that could be improved. In the final round, the team identified searches that could be considered a failure (where material judged likely to be mostly irrelevant filled at least the first 10 results), and made judgment calls about which configuration delivered the basic sets. In most cases, the judgments were clear-cut. Where the team disagreed, there was a minimal attempt to reach consensus, however these situations often revealed the broader problems with judging relevance as the team were guessing what different users might want, and we decided in most cases to note the differences rather than come to a decision over what was "better". Of the 50 test searches trialled, only one was considered "failed" in the final iteration of the system that has gone into production.

We were also able to trial the impact of including and excluding different MARC fields and developed a list of the most useful MARC fields to use to represent relevance such as author, subject and title. The final list of MARC codes being indexed was much tighter than the list we started with. A bibliographic record is structured to provide as much information as possible to the user, but this is not always an advantage in relevance ranking. Often records for items that are derivative of another item contain much more information than the original, which tends to mean they rise to the top of relevance ranked lists. For example, publications put out by the Sydney Morning Herald, often have the term Sydney Morning Herald in the main author, statement of responsibility, and additional title fields. This is also a problem with translations.

Through this testing, we were able to refine a relevance ranking system, which was implemented in the November release of Libraries Australia. We are also continuing to work on a prototype search system, which is underpinned by different search software, to investigate both further refining the relevance ranking, and implementing other search services such as clustering, and recommendations to refine the user's search. This prototype is available for public testing and comment at http://www.ll01.nla.gov.au/.

During the course of this project, some of the people consulted felt the addition of easy fielded searching to the basic search would be more valuable than relevance ranking, while others acknowledged that relevance ranking might be useful for novice users but commented that they would still use the advanced search to retain control of the search results. This reflects that many believe relevance ranking and fielded search are mutually exclusive approaches to resource discovery. However, relevance ranked results can still improve a fielded search (for example, an author search on Patrick White will still return works by many different authors). At the same time, relevance ranking does mean that some searches currently carried out as fielded searches may be just as efficient as a simple key word search. In the same way, offering clustering, limit check-boxes and other features are a different way of allowing users to make the same decisions that an experienced searcher makes when limiting a search to specific fields or using boolean operators and proximity searching. It will be interesting to see what impact good relevance ranking makes on the search patterns of experienced users over time.

The future: Clustering

Relevance ranking will improve the search experience for most users, but it is not enough to ensure that users can find what they want in large result sets. Firstly, users express their information needs very differently to each other. Two users entering the same query are likely to have very different expectations from the result set. If we managed to

configure relevance ranking to exactly suit one user, it would almost certainly be much less useful for the rest of the population. Secondly, many users still use search queries so broad that the result set is unlikely to meet their information need.

Search logs from Libraries Australia illustrate this point sharply. While it may seem apparent what some users are searching for from their query ("topless girls"), with others, it is impossible to guess ("uranium", "population", "terrorism", "modern history" are good examples.)

Rather than simply expecting users to try another search, however, emerging search technologies provide users with more options – to choose from "clusters" taken from their search results, to help refine broad queries, to change to a related or broader search term where there are few results returned, or to follow paths that others looking at similar material followed. There are a range of new technologies that we are looking to incorporate into Libraries Australia in the medium term. One of these is clustering of search results.

A good example of how clusters could be useful would be for the search "civil war". In a sidebar, users could be offered access to groups formed from within the result set, enabling users to go directly to the Spanish, American or English civil wars, for example, or to resources that deal with civil wars in general. Users might also be offered material clustered by type: personal accounts, fiction, textbooks, journal articles etc. Clustering resolves some of the problems inherent in keyword searches. When a user types a name into the search box, do they want works by that person or about them? If they search for DVDs, do they want material about the format or in that format? With clustered results, we can let the user choose.

Several libraries worldwide are working on implementing clustering. Two very different approaches are at the North Carolina State University Library, and the Aquabrowser project. NCSU has a highly structured approached to clustering, based on the same MARC fields as the Lucene prototype, but with the results much more tightly organised. For example, this enables a good ability to "surf" through subjects, but also necessitates a cluttered interface presenting the user with dozens of options. A cleaner interface based on similar structures, with a drill down approach, is presented by OCLC's new Worldcat.org service. The Aquabrowser attempts to present a graphical mind map of the associated terms, with the option for users to click on related terms to take them to another search, or a refined set of results.

The clustering currently implemented in the Library Labs prototype is based on metadata culled primarily from the subject entries, the leader and 008 fields in a MARC record. It offers users the opportunity to refine their results by subject, by Conspectus discipline, Dewey Decimal Classification, and by the format, literary form or nature of a work.

However, particularly when it comes to subject clustering, it is useful to have a thesaurus, taxonomy or ontology that defines the relationship between different subjects. This enables us to not only group together items with the same terms in the same fields, but to suggest to users how to access related subjects, drill down to more specific subjects, or broaden the search to more general subjects. The Library Labs prototype has recently been enhanced to offer some basic suggested further searches, based on LCSH authorities. However, this is in the very early stages, and has already shown that it will take some work to use LCSH structures most effectively. Much more investigation is

needed to refine a proposal that could be implemented. There is a considerable international discussion about whether Library of Congress Subject Headings are the best framework for such search tools, and if it is possible to develop an alternative, and this discussion is progressing very quickly. Some of this is reflected in the appendices to the report prepared by Karen Calhoun for the Library of Congress [5]. The LibraryThing service has been able to build a useful folksonomy based on user-generated tags. The strong success of this project offers up ideas for other potential sources of data.

Any large catalogue will describe different versions of the same work. Grouping these together using the model described by the Functional Requirements for Bibliographic Records report[6] is another type of clustering which we've just begun to explore.

Recommendations

One of the biggest reasons for the success of search systems such as Google and Amazon is their incorporation of users' patterns of behaviour into recommendations. Google uses the page rank system to work out what connections people draw between web pages, and what they cite, and factors that into its relevance ranking. Amazon uses user ratings, user-created lists, searching and buying patterns to offer users a range of recommendations.

These functions have been enormously popular, as they enable users to follow paths created by common experiences, not by the judgements of an authority author, or of a single cataloguer. They have helped develop a new form of browsing, where users explore options generated by the actions of their peers.

Because they are based on the actions of others meeting a need, they tend to offer a quick way around some of the quirks built in to authority-based systems. (For example, a search in Amazon on "Huck Finn" quickly returns a note explaining that most users who search for Huck Finn ended up buying the Adventures of Huckleberry Finn, even though this is not returned in such a search).

Systems can simply identify the most popular items in a result set and allow the user to access those. This data could come from a number of sources: holdings data, records of previous search patterns, circulation data or user-inputted data such as ranking of particular items. Local data could be supplemented from data sourced externally, such as Amazon rating, or LibraryThing tags.

The data could be presented in a number of ways: a "most borrowed items in your results" list, or a "most highly ranked" items list, or even a simple list of the most widely held items.

The second approach is to try to recommend items that previous users linked in some way to the search term entered. This could incorporate data such as tags added by users to the item or information gathered from previous search patterns. It could produce results such as a "users who searched for x ended up borrowing y" feature, or a "common tags identified with your search" feature. We could even enable users to create their own lists or subject guides, which could be returned in a sidebar. All of these are potential future developments to explore in discovery and access services like Libraries Australia.

Tagging and social networking

Tagging, where users of a system can attach "tags" (generally keywords) to an item's record, is a growing feature in web-based bibliographic systems. It provides a way for users to enhance the information available in the catalogue, and new ways of linking related items, and developing scholarly communities.

Tagging is a rapidly expanding phenomena, and there are now several examples of tagging in bibliographic systems, including LibraryThing, and the recent integration of tags into Pennsylvania State University catalogue. The Connotea project is also worth looking at, as of course are Flickr and del.icio.us. LibraryThing now boasts seven million user-added tags.

The potential of this technology is still unfolding. In LibraryThing, and in Connotea, different users apply tags for different purposes. Some add information that could have been added by a cataloguer – subject or genre information. Others use tags to identify a personal use for an item, identifying it as a university text, or as light reading. These tags can be searched as part of the record, incorporated into clustering and other schemes. They provide an entry point for users searching on terms that are not commonly applied as subject headings, such as "manga" or "zine", or for ways that books are not catalogued, such as "holiday reading" or "literary classic". Because it is possible to measure how many users have applied a particular tag, ranking can rely on what a large number of people thought, rather than an individual's judgement.

One of the most interesting developments for tagging, is the facilitation of communication and the creation of new communities. Sometimes, tags are used to create lists – such as denoting undergraduate texts for a subject, for example. Other lists provide resources for general users, such as subject guides, or "Top 10" lists. Users in LibraryThing and del.icio.us can look at others' tagging decisions, and follow those to new resources. At the University of Pennsylvania, PennTags users can comment on each others' tagging decisions, enter descriptions of and create projects that organise and share their research sources (catalogue items, journal articles and web sites).

Taking our search to the users

Fewer information seekers now start their search at the library at all – the OCLC's recent study of college students put the proportion at 24% (14% at the physical library and 10% at the Library's website) [7 p. 11]. In part this is because many information needs can be met elsewhere - many reference queries that would have once necessitated a visit or a phone call to a library can now be answered through a quick visit to Google or Wikipedia. This is not the whole picture, however. The OCLC study also found that more than half those surveyed considered library materials would be useful for their search .

So why would they not come to the library first? Suddenly, big is better. Because relevance ranking means that large result sets are not a problem, more users want to simultaneously search large collections of material, which may include library material, but is not restricted to it. Many users start a broad search on the Internet, or a favourite web service, and then slowly drill down to more detailed information through browsing of many different resources. Services such as A9 offer the ability to simultaneously search many different resources, including monographs, images and reference services.

University metasearch portals also provide access to an increasingly broad range of resources. This means that there are thousands of potential Australian National Bibliographic Database users who will never come to the Libraries Australia website directly. The AND, which underpins the Libraries Australia service, contains a wealth of items that would meet many of these users' information needs, particularly the rarer and unique items unlikely to be found elsewhere. This necessitates bringing the ANBD to users, rather than the other way around.

The library has already made the ANBD available as a target for systems such as libraries and university metasearch portals through the Z39.50 search and retrieval protocol. In practice, the library's preferred metadata repository and search system will need to support a range of protocols to enable it to be accessible to different communities. Z39.50, while a legacy standard, is still widely used by the library community and will need to be supported in the medium term. SRW/SRU, its designated replacement, will enable searches of the ANBD to be deployed on any web page using a simple web service.

Support for OpenSearch was released in November 2006. OpenSearch is a simple protocol supporting very basic single-field searching and the formatting of search results in an easily machine-processable format which promotes their reuse and aggregation. Supporting OpenSearch means that users can search Libraries Australia directly from services using it, including from within Firefox 2.0 browsers once it is loaded as a search engine in the search toolbar and the A9.com website. In order to meet these diverse requirements and to increase access, Libraries Australia will enable support for multiple search and delivery protocols. In many online spaces, simple keyword searching and server choice will be the primary search option. This makes it imperative for the ANBD target service to deliver relevance-ranked results in these environments.

The National Library of Australia is also involved in developing other federated metadata repositories and services to allow users to search for specific materials across a vast collection, or series of collections. Good examples of such projects are People Australia, Picture Australia and Music Australia. Libraries Australia is strengthened by providing access to the resources in these repositories as well. These resources should be available to users of all search interfaces.

Internationally, there is an emerging discussion about the relationship between architectures built upon a meta-search basis, where a single interface provides a search across many targets, and architectures built upon aggregating a large number of resources into a single target. While the former may offer an easier framework for collaborative projects and flexible, the latter approach provides an easier framework to deliver new search features such as relevance ranking, clustering and recommendation services. This has led some libraries, most notably the University of California library, to decide to consolidate as much as possible into a single data store. There should be no dumbing down of search results just because two targets have to be searched. There are challenges here for relevance ranking that might be met by ensuring that each target delivers a separate meaningful cluster of records, for example, pictures, books or articles.

Conclusion

The current time is one of expanding information access, and within this, there is a key role for libraries if we are ready to take up the challenge. If we have confidence in the

continuing relevance of our collections, then we should have the confidence to make them more accessible, to let people know what we have got. In this period, Libraries Australia is one of Australia's key assets, providing a gateway to almost 42 million items held by Australian libraries across the full range of human knowledge: some highly entertaining, some of rare value and great worth. We now have the tools available to make these resources much easier to find. If we can harness the power of Libraries Australia as an end-user search system, we will be playing a significant role in making Australia a smarter, more-informed society.

1. Australian Bureau of Statistics, *Household use of information technology, Australia, 2004-05*. 2005.

2. Anderson, C., *The long tail*, in *Wired*. 2005.

3. Dempsey, L., *Libraries and the long tail.* D-Lib Magazine, 2006. **12**(3).

4. Bates, M.J., *Task Force Recommendation 2.3 Research and Design Review: Improving user access to library catalog and portal information*, in *Library of Congress Bicentennial Conference on Bibliographic Control for the New Millenium*. 2003: Washington.

5. Calhoun, K., *The Changing Nature of the Catalogue and it's Integration with Other Discovery Tools*. 2006, Library of Congress: Ithaca.

6. Bibliographic, I.S.G.o.t.F.R.f. and Records, *Functional requirements for bibliographic records : final report*. 1998, International Federation of Library Associations and Institutions.: Munchen.

7. De Rosa, C., et al., *College Students' Perceptions: Libraries & Information Resources*. 2006, OCLC Online Computer Library Center.