Supporting User Search for Discovering Collections of Interest

Buchanan, George
g.r.buchanan@gmail.com
School of Informatics, City University, London

Dodd, Helen
cshelen@swansea.ac.uk
Future Interaction Technology Group, Swansea University, Swansea

Humanities researchers often draw their information from a diverse set of sources, which are unlikely to be found in one digital library or other collection. Whilst a considerable volume of technical work has been undertaken to unify a number of digital libraries into one whole - "federated" or "distributed" digital libraries - the real-world adoption of this technique is riddled with conceptual problems and organisational practicalities. In consequence, there is little likelihood of there arising one "super-gateway" to which an active researcher in the humanities can turn for the whole of their information seeking. Indeed, even in the sciences, this idealised situation is relatively rare. As a result, a humanities researcher will often need to identify a number of different collections that serve their regular information needs well [Buchanan *et al.* 2005]. Each of these collections would be likely to provide useful literature relevant to their field of study. However, constructing this list is problematic. At present a user can only generate such a list of 'good' collections from months and years of incidental discovery and recommendation.

In the past, this need has been addressed in a number of different ways. The Humbul humanities hub - later part of Intute - was started as a human-maintained list of online resources for the humanities and arts. It relied upon hand-crafted entries created by researchers (often postgraduate students) that described each collection and suggested its potential uses. One positive advantage of this method is that it develops a single list of many collections, and is open-ended. However, there are problems that emerge when a user tries to find collections within the hub. This approach necessarily requires the creator of an entry to double-guess the likely tasks of another user and a substantial cumulative effort over many years. It is resource expensive in terms of creation, difficult to maintain, and an entry is very unlikely to be able to cover all likely uses in exactly the terms that another researcher may use. Whilst a positive benefit is that the system provides a 'human readable' overview of each site, if a user provides a search to the system, it will only use the relatively small amount of material entered by the researcher who entered the site onto the system. Compared to the information available within a single library, a brief descriptive entry is prone to have insufficient information, and may overlook the searcher's interest altogether, or provide a disproportionate representation - greater or lesser - of the volume of material relevant to their work.

In contrast, highly technical approaches have been undertaken to create a central, canonical resource through "metasearch" techniques: where a central service endeavours to offer a synthetic unification of a number of DL systems [Thomas and Hawking 2009]. In these methods, the user provides a sample query that is then automatically sent to every individual digital library. The metasearch engine combines the results from the separate libraries into one whole, and returns the unified set of results to the user.

One limitation to the metasearch approach is that the list of libraries or search engines supported is usually fixed. Metasearch on the web often relies on "reverse-engineering" the HTML output from each single search facility that the metasearch system uses. This requires extensive maintenance work, and lists of available collections are thus often fixed.

Another concern with this approach is that from a conceptual view, it is vulnerable to poor understanding or even utter ignorance of the operation of each constituent DL. Different DLs may interpret the same search in radically different ways. Another issue is that as these services normally send a search to each of their constituent DLs for each search given by a user, and this means a substantial overload of search activity for each constituent DL. Practically, this is clearly ineffective and costly in

terms of computation and, ultimately, hardware resources.

One method proposed to minimise such waste is to provide a "database selection" algorithm that pre-selects only the better DLs to search, and these are then queried by the metasearch system automatically [Thomas and Hawking 2009, French *et al.* 1999]. The metasearch system then combines the result sets from each chosen DL and provides a single ranked list of matching documents. However, the same problems with general metasearch return: how results are combined into one remains an issue, and users are known to be poor at reconstructing the best sources (i.e. the sites with the largest volume of relevant material) from such lists.

In our research, we do not attempt to circumvent the conceptual problems of unifying search result lists from different libraries with varying matching algorithms and heterogenous vocabularies. Rather, we aim to embrace the diversity of information sources, and suggest to a user the libraries that are more likely to contain good-quality information for a given query. In this regard, our approach is similar to meta-search. However, there are key differences in our method: first, we provide the user with a list of likely "best" DLs, rather than individual documents; second, we do not attempt any merging of result lists or other conceptually fraught manipulation of retrieved material; thirdly, use DL technologies such as the OAI protocol for metadata harvesting (OAI-PMH) to provide an open-ended list of target collections [Van de Sompel *et al.* 2004], in contrast with the fixed list limitation that is commonplace in metasearch; fourthly, we are reconsidering the best matching algorithms from a user-centred perspective, rather than from currently commonplace information-retrieval based metrics that may poorly match a user's requirements when their aim is to discover rich, large-scale sources of information on a particular topic.

Our current research has probed this fourth and final issue. What we have uncovered is that many of the current database selection algorithms privilege two sorts of collections: first, large collections; second, collections in which a sought-for term is rare. There are good reasons why this is appropriate, from a technical information-retrieval point of view. For example, a term that provides strong discrimination within a given collection is likely to produce a clear, consistent list of matching documents. However, from the perspective of a humanities researcher investigating a new topic, and who is seeking libraries with good coverage of the topic, these extant IR measures seem a poor match against their requirements. A digital library that contains a high proportion of documents with the sought-for topic is arguably very likely to be of long-term value to them. However, this pattern is exactly the opposite to that desired by the existing database selection algorithms, where a strongly distinctive term - i.e. one that matches only a few documents - is ideal. Similarly, absolute size is not necessarily a criterion. Often, a small but specialised high-quality collection is of critical value in orienting the humanist in a new domain. Furthermore, our own previous research suggests that humanities researchers do not trust computer-based models of relevance: they prefer computer systems to err slightly on the side of deferring such decisions to the user.

In assessing the current state-of-the-art, we have developed an experimental apparatus that permits the testing of several algorithms in parallel. Conceptual, idealised scenarios can be test-run, and then the same tests applied to sets of collections gathered by the researcher to retest the same scenario on real data. This permits us to assess any algorithm against nominal and real data, and against a number of alternatives. This test environment has already demonstrated that many current database selection algorithms perform very poorly against the ideal criteria for the task that we seek to support, and that even the best are far from optimal. The bias towards large collections outlined above has been reiterated in practice, and we have also uncovered the problem that the numerical ratings produced by good algorithms follow unhelpful patterns.

One frequent problem is that subtly different scores for different collections against a particular search can be produced from profoundly different underlying coverage of search terms in the collections. This underlying problem is manifested in different ways. Common oddities include relatively high scores being given when only one term is matched

(albeit many times) and ‘normalisation’ of scores meaning that very different matches between collection and query result in marginally different final scores. This second problem means that common methods for deciding on a list of ‘best’ matches do not work, and it is difficult to decide the criteria to use to interpret a score into a final recommendation.

A considerable body of further work is required. Whilst we are confident, from the current literature, that our current results are, for idealised scenarios, closer to what is required, the problem that we seek to answer is as yet poorly understood. We need to investigate further not only the technology, but also the human context in which it will operate, and through this develop a more sophisticated and accurate model of what humanities researchers would ideally require as the output of our system.

References

Buchanan, George, Cunningham, Sally Jo, Blandford, Ann, Rimmer, Jon, Warwick, Claire (2005). 'Information Seeking by Humanities Scholars'. *9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005).* Vienna (Austria), September 18-23, 2005, pp. 218-229.

Thomas, Paul, Hawking, David (2009). 'Server selection methods in personal metasearch: a comparative empirical study'. *Information Retrieval.* **5**: 581-604.

French, James C., Powell, Alison L., Callan, Jamie, Viles, Charles L., Emmitt, Travis, Prey, Kevin J., Mou, You Y. (1999). 'Comparing the performance of database selection algorithms'. *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '99).* Berkeley, California, United States, August 15-19, 1999. New York, NY: ACM, pp. 238-245.

Van de Sompel, Herbert, Nelson, Michael L., Lagoze, Carl, Warner, Simeon (2004). 'Resource Harvesting within the OAI-PMH Framework'. *D-Lib Magazine.* **12**. http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html.