On the relation of search and engines

Frances Johnson
The Institute of Information Research (TIRI)
Manchester Metropolitan University
f.johnson@mmu.ac.uk

The Information Retrieval course can focus its teaching on the data structures and algorithms of the systems or on understanding the user interactions and the social contexts of IR, depending on the target audience. This paper explores the relation between the core subjects of Information Seeking Behaviour and Search Engines (the retrieval technology) and, with regards to curriculum development, examines the extent to which the teaching of one can help understand the crux of the issues in the other. The notion developed is that teaching IR technology in the interpretative context of understanding search puts search on the curriculum and provides a depth to understanding the challenge for search in modern information environments.

Information Retrieval, Search engines, Information Seeking Behaviour, Search, Curriculum Development

1. INTRODUCTION

Defining an IR module is complex because IR is affected by developments in the environments in which we create, represent and access information. Jones (at TLIR, 2007) addressed this in developing a problem-based module pointing out that the latest applications and topics only emphasise the need to focus on the basic principles. This paper begins to explore a possible synergy in the teaching of the core topics in Information Seeking Behaviour (ISB) and Information Retrieval (IR) and sets out to explain how the teaching of IR technology can enhance understanding of the skills involved in conducting search in the modern information environment. Over the past decade the ISB and IR research communities have been brought together for the development and evaluation of systems which support the user engaged in the process of information seeking. In this paper the reciprocal relation of ISB and IR is considered, focusing on the users' skill in carrying out a search, what is involved and how should it be taught. The proposition presented is that teaching IR technology helps develop an understanding of search, a sub activity of ISB, such that it can be taught and honed informed by knowledge of the supporting systems.

1.1 IR in the curriculum

IR in the curriculum for Computer Science (CS) provides an interesting application for the teaching of computing principles, such as data structures and programming algorithms. Bawden (2007) outlines a core module for IR in the Library and Information Science (LIS) curriculum which focuses on the organisation of information within established contexts and social purposes. At a glance the programmes appear similar covering the core topics of indexing and search based on Boolean, statistical and web retrieval models, with the study of ISB and evaluation providing the notion that IR takes place in challenging contexts. Yet, with regards to the computational detail, it may appear that there is no discernable relation between the two programmes. The IR module in CS covers the technical detail of data structures, models and algorithms of IR. In LIS depth of coverage tends to be found in the contexts in which IR takes place, the user interaction, the interface and usability, rather than with the back end. This, however, presents the question – how much computational detail should be given in the LIS programme and specifically at what level is the teaching of IR technology meaningful to the LIS student? In the CS programme, the study of ISB provides an important context giving depth of understanding to the challenges set for the development of IR systems. ISB - the information need, the uncertainty of relevance, and with the potential ambiguity of language used to express the query - together define IR and explain why effective IR is difficult. In turn, the LIS programme can provide an interpretation of the research and development coming out of the CS community. It is this interpretation which can determine the extent of the teaching about IR technology. The isolated one or two sessions covering retrieval models can leave the students out in the cold thinking that IR is hard and is not really of much concern to them. The solution presented here is that the study of IR technology should be integrated into the programme which seeks to enhance students' study of ISB and specifically pitched to develop understanding of search behaviour. That is, the reciprocal to the study of ISB in CS – the study of IR technology in LIS programmes provides depth of understanding of the challenge of search.

2. TEACHING SEARCH

On the development of the digital library curriculum Saracevic (2001) posed the questions why teach, what and how. We can also ask these questions with regards to search – why teach search, what teach about search and

how teach about search? The challenge of IR from the user perspective is how to search and Belkin's (2000) description of the challenge gets to the crux of the matter "How to guess what words to use for the query that will adequately represent the person's problem and be the same as those used by the system in its representation". Yet this description is deceptively simple. What lies behind this description is perhaps best understood through the study of the indexing and search techniques and technology of the systems supporting the users' search.

2.1 Why teach search

The immediate response to the question *why* teach search is to point to the role of the search intermediary in the profession. Whilst there has been an ongoing shift to end user searching, most LIS IR courses will provide the training to hone students' search skills and some graduates continue to take up posts to conduct search on a professional basis. Typically the training involves the formulation of search strategies implemented on host services, such as Dialog, that market themselves as tools for professionals provided by professionals. The question, *why* teach search? is perhaps more perplexing when asked in mind of the students who regard their work cut out in the modern search environment and, accustomed to making Google their first port of call, typically regard search as a simple activity. For many intents and purposes it is; the prescriptive IR technology or engine, powered by techniques such as link popularity, collaboration and personalisation, is undeniably effective at locating the information we want. Yet the technology which attempts to infer users' intent, it can be argued, does so to the detriment of the searchers' ability to devise a search strategy. Searchers assume that search engine technologies can and have found them the best information, unaware of the way their results are found and of what else might be there. This user is "satisficed" (Griffiths et al 2007) – stopping when they have achieved some relevant results rather than seeking to optimise the precision that they have already achieved. This is partly what is railed against in the literature which in Brabazon's (2004) words suggests Google as the white bread for the mind. Thus returning to the question, *why*, the implication is that, if not, searchers will become lazy bought about by computational offloading (the shifting of intellectual effort from the user to the engine). Again in defence of the prescriptive engine, least effort is an important criteria of success but as is effectiveness. If the modern search engine diminishes the intellectual effort required from the user and/or disengages them from some search strategy (probably best observed in users' failure to recover from a failed search or to recognise relevant results when successfully retrieved) this in itself poses the question of what should we teach about search and how.

2.2 What to teach and how

The IR module in LIS courses, pre-web, taught the student how to search with regards to the skill and expertise expected of the search intermediary. Today a proportion of the module is likely to concentrate on the teaching of 'online searching'. With regards to what and how we should teach search it is interesting to note that the skills of the search intermediary are most effectively taught through knowledge of the architecture or workings of the retrieval system. Reverse offloading perhaps – when knowledge of the extent of system's capabilities (how it interprets the query to retrieve relevant items) indicates the intellectual input required from the user. Students can be introduced to the online database as the computerised card catalogue with increased access points from full text and a range of search features which enable the searcher to deal with the vagaries in language when formulating the query. The search intermediary can develop search expertise though an in-depth understanding of the processing steps involved in creating the bibliographic database. Knowledge of how the inverted index is created, the steps of tokenisation, stop word removal, stemming algorithms, parsing for indexing at word and phrase level, the splitting of the posting and dictionary files for binary search all give the searcher insight into the functionality of the search features such as the proximity operators, field limiting and the application of Boolean logic and set creation in query formulation. Critically, with regards to the question *what to teach* about search, it is by teaching search based on detailed knowledge of the system backend that emphasis can be placed on developing the searcher's skill in query formulation and strategy for the manipulation of the search towards the desired outcome. A strategic approach can be practised based on a search plan with terms identified for the key concepts of the search. System search features and search strategies can then be used to group the search terms, broaden or narrow the sets and combine the search lines to the desired outcome. The searcher has a certain control over the queries formulated to match against the system's representations and the search is a problem solving process with assessments made of the relevance of the retrieved items to conclude the search or obtain feedback. Drawing on conceptual knowledge to accomplish the task, the system can be understood to be the tool while the searcher concentrates on finding the query terms to represent the information need and to match the representation of relevant items stored in the system.

2.2.1 Teaching search strategy in web environment

The modern search engine, however, does not (obviously) present the user with features to formulate a query, arguably disengaging the searcher from the search process with the expectation of immediate satisfaction. The search engine, without the search history box or the search features for the user to input a formulated search,

provides little to encourage or support search as an intellectual, skilled and/or strategic process. Unless we expect students to take what they have learnt about searching on the traditional online database to the modern search engine (or indeed other IR systems such as the digital library) there does not appear to be much we can teach about search on the modern statistically based engine. This is certainly the case if we think about search as a procedural activity, but as a conceptual process (as seen in the search intermediary working with the bibliographic database) with the use of interpretative skills that concentrate the mind on the query analytically it would appear that the modern engine, likewise, can teach us quite a bit about search. If search is best understood by its practitioners to be an interaction between the user and the information represented in the system it helps to have a view of the system's text processing in delivering its contribution to the communication/process. Learning about the retrieval models on which modern retrieval systems are based may provide the searcher with a mental model of retrieval which promotes an analytical approach to the search. The vector based model of modern retrieval systems presents a sense of search as an activity which focuses on the challenge to find the words used to capture the nuance of a query, to extract cues or words as directional signposts and manipulate those terms with respect to the items represented in the system. Understanding the system processing involves learning about its weighting of the index terms, based on calculations of word frequency and relative frequency (*tf x idf* scores), and representation of the document in a n-dimensional space for the calculation of document-to-query similarity. Further detail on techniques used in clustering and relevance feedback, when taught in this context, can provide a sense of IR as an iterative process in which the system groups documents by similarity and indeed dis-similarity to the query. Tenopir (2001) stated her informed opinion that Dialog should still be taught in post-web times because the command driven interface shows the searcher how the system works and allows the searcher to control the search and its implementation. By extension, or generalisation, it could be that teaching how retrieval systems (including search engines) work enables the student to understand what it is to search: what it means to select good query terms representative of the information need and with the power to discriminate relevant from non-relevant, and, as an interaction, the gains to be had in working with the system feedback to refine the query and its outcome. Understanding the system's processing of the query helps gain an understanding of the intellectual aspect of search from the user's perspective and whilst there is no hard evidence, general feedback from students indicate that they feel they have a far greater understanding of and confidence in implementing an effective search on a retrieval system.

3. CONCLUSION

We started out by briefly considering the relation of ISB and IR in the taught modules found in CS and LIS programmes. ISB clearly provides a context for understanding why IR is a challenge and can help gain a view of information search as a complex activity often initiated as a result of having some information need and with aspects of that experience (such as making relevance judgements) affecting the interaction (in formulating and evolving the query). The notion developed here is that the study of the IR technology can, in turn, help students to understand the challenge of search from the user perspective. Teaching search is not straightforward – it is a process which is affected by context (user and task): yet at the same time it can be described quite simply as a matter of finding of the right query terms. A key to understanding search lies in its development as a skill in a module which covers in some detail the architecture and models of information retrieval systems. The inverted index as a searchable data structure, the ranking and clustering of the statistical engine and even the inference of intent of the web engine provide insight into what is involved in conducting a search on these systems in the quest for information. In conclusion it is argued that IR technology should be taught in detail on the LIS programme and most effectively in the interpretative context of understanding and developing search skills. This is not to suggest that new topics are introduced into the programme, rather that the interpretative context is built up through the coursework and the questions posed. This paper merely sets out the proposition which provides the context for the module and its learning outcomes upon which readings, examples, exercises and questions are based around. This exploration of the relation of users' search and the search engine technology enables the IR course to cover the key principles and practices of the field and furthermore provides a basis on which to apply this knowledge to emerging key topics (such as search interface and information visualisation) related to core concerns in the implementation and integration of search in modern information environments.

REFERENCES

- [1] Bawden, D. Vilar P., Bates J., Steinerov J. and Vakkari P. (2007) Information retrieval curricula; contexts and perspectives. *First International Workshop on Teaching and Learning of Information Retrieval (TLIR'07)*, London, January 2007, [eWIC](http://ewic.bcs.org) (<http://ewic.bcs.org>)
- [2] Belkin N.J., (2000). Helping people find what they don't know. *Communications of the Association of Computing Machinery*. p.58-61.
- [3] Brabrazen, T. (2004) BA(Google): Graduating to information literacy. *Keynote paper at the IDATER on-line conference* (Loughborough University, August 2004).

- [4] Griffiths, J.R., Johnson, F.C. and Hartley, R.J. (2007) User satisfaction as a measure of system performance. *Journal of Librarianship and Information Science*, 39, 142-152
- [5] Jones, G. F. J. (2007) Teaching Information Retrieval Using Research Questions to Encourage Creativity and Assess Understanding. *First International Workshop on Teaching and Learning of Information Retrieval*. London
- [6] Saracevic, T. & Dalbello, M. (2001). A survey of digital library education. *Proceedings of the American Society for Information Science*, 38, 209-223.
- [7] Tenopir, C. (2001). Why I Still Teach Dialog. *Library Journal* , 126 (May 1, 2001): 36-38.