Teaching Information Retrieval Using Research Questions to Encourage Creativity and Assess Understanding

Gareth J. F. Jones
School of Computing, Dublin City University
Dublin 9, Ireland
Gareth.Jones@computing.dcu.ie

Abstract

The study of information retrieval has increased in interest and importance with the explosive growth of online information in recent years. Learning about information retrieval within formal courses of study enables users of search engines to use them more knowledgeably and effectively, while providing the starting point for the explorations of new researchers into novel search technologies. The nature of information retrieval as a topic also makes it an ideal subject for developing a range of interdisciplinary and transferrable skills in those studying it.

Keywords: information retrieval, teaching and learning, problem analysis and solution, student assessment

1. INTRODUCTION

Interest in information retrieval (IR) as a subject of study has increased significantly in recent years. This has been driven by the very rapid growth in largely unstructured online information repositories, principally the internet, but increasingly other digital media sources such as audio and video, and the need to be able to efficiently locate information relevant to a user's information need from within these collections. The new challenges and opportunities for IR technologies arising from varying user needs, expectations and expertise, and available information sources mean that IR has emerged as a dynamic and rapidly developing subject with a vibrant and growing research community. Despite the proliferation of new areas of IR research such as bioinformatics, question answering, multimedia IR, topic tracking and web search, the fundamental issue of IR remains satisfying the information need of a user expressed through some form of search request.

In order to fully appreciate, exploit and contribute to advances in search technologies students studying IR need to acquire a sound knowledge of the fundamental issues and techniques, open questions and relevant research strategies. Learning about IR methods enables users of search engines to use them more knowledgeably and effectively, while providing the starting point for the studies of new researchers. The nature of IR as a topic also makes it an ideal subject for developing a range of interdisciplinary and transferrable skills in those studying it. In many ways IR is no different to other subjects, however I would argue that the principles of IR are sufficiently easy to grasp, and in combination with the various practical challenges of information searching, IR forms an ideal subject for exploring creative teaching, learning and assessment methods.

In this paper I review some of my experiences in developing, delivering and assessing modules in IR. These are related to studies in student learning, expectations and assessment, and my observations on student prerequisites and the potential for original research contributions in student projects. Section 2 considers the content of an undergraduate module in IR, Section 3 reviews some relevant studies in student learning, Section 4 describes the design and outcomes of a final year BSc module in Information Access that I developed some years ago, and Section 5 concludes with details of outputs from this module and outlines details of a more advanced module in incorporating IR.

2. WHAT SHOULD BE COVERED BY A MODULE IN INFORMATION RETRIEVAL?

Information retrieval is typically taken as a one semester module within final year undergraduate and masters programmes. Students may of course have studied related topics in other modules, such as multimedia technologies, the internet or computer networks in general, or one or more topics in natural language processing. However, such related modules are often also optional and knowledge of these topics cannot be assumed of the whole cohort opting to take IR. Within a programme of study where these topics may have been are covered

elsewhere by some students, the IR module must necessarily repeat some material, but must seek to present and address it in an alternative way relevant to IR.
More fundamentally, since IR is only a single semester module, it is in my view unrealistic to seek to teach both the basic concepts and give detailed coverage of a wide range of the related techniques and algorithms. In my experience, it is all too easy for instructors to adopt, probably without considering the implications, a strategy of determining that students must be exposed to a very wide range of material in great detail, since all this material is "vital" to an acceptable working knowledge of the subject. Or course, the danger of doing this is that the module becomes so packed with content that in order to cover everything the instructor finds it necessary to deliver it as fast as realistically possible in classical lecture delivery mode where students are expected to sit and absorb as many facts about contemporary IR methods as possible. If one steps back to consider this scenario for a moment, it is clear that this loses sight of both the objectives of effective teaching and learning from the students' perspective [1], and also risks the students not being able to see the wood the trees in terms of appreciating the fundamental issues of IR amidst the details of current techniques which may quickly be replaced within a year or two.
Taking these issues together, some years ago while working at the University of Exeter, I developed a module called *Information Access*. Before describing the design and methodology of the Information Access module, the next section briefly outlines some of its underlying motivations based on research in student learning. Much of this review is taken from [1] which provides an excellent summary of student learning taken from a wide range of sources.

3. TEACHING AND LEARNING METHODS AND INFORMATION RETRIEVAL

3.1. Student Learning Modes and Assessment

Student learning can be generally classified into two forms: *surface* (or shallow) and *deep*. Students engaging in surface learning are generally found to be attempting to store information without much thought. Learning is seen as a process of acquiring facts related to a subject, and learning the principles and procedures associated with them. The student sees the role of the instructor as being to present information and the students' role to reproduce this information in an examination in order to demonstrate that they know it. This is the classical lecture presentation process involving large amounts of rote learning on the part of the student, and fairly unchallenging lectures from the perspective of both the lecturer who presents a prepared lecture "script" and the student who is expected to listen to (and it is generally assumed "understand") the material without actually doing anything else. By contrast, in deep learning the emphasis is much more on thought rather than memory. Rather than learning by rote the "words of wisdom" from their expert lecturer, students must actively integrate new ideas with those already possessed. Students using a deep approach look for the fundamental principles associated with the subject. They distinguish the principles of the subject from examples which demonstrate these principles in applications and are then able to exercise imagination within the subject.
Classical lectures often encourage the surface approach and actually discourage the very intellectual skills of thinking, integration and imagination that higher education claims to foster. It is assumed that students will reflect on material delivered in lectures and explore material afterwards in private study. But high teaching and course assignment loads and, let's face it, laziness on the part of the student, coupled with associated module assessments and examinations which are often driven by factual regurgitation in examinations make matters worse.
It is often observed that interest increases and learning is improved if students are asked questions rather than told facts. Delivering lectures in this form is harder work for the lecturer who must be prepared to do more than recite information and be prepared to challenge and engage students, and for the students who must really listen, engage and think during the class.
It is frequently said that students are motivated by assessment. Thus if they can see the relevance of the material to the module assessment, students are more likely to be motivated to engage with it. The desire to achieve generally improves motivation and learning, but students need to know what is to be achieved! Thus it is important to tell them the objectives of the module and each lecture at the start, so that they know what they should learn from it. In terms of assessment, one can rely on reciting information, but if the students have actively engaged with the material in classes and have been encouraged to approach the material from an imaginative and creative perspective, this can be pursued in the assessment as well. This leads to the opportunity for the examiner to ask questions which explore the candidates understanding of the principles of the subject and apply it to novel problems, rather than reciting lecture notes or filling in the gaps in small variations in examples taken from lectures.

3.2. Teaching and Learning Information Retrieval

Information retrieval is a subject built around fundamental principles which are generally accessible to students of a wide range of skills and abilities. The rapidly developing range of technologies associated with IR mean also that it is a subject in which imagination plays a key role in new developments. IR is thus a great subject for students to

learn since it enables them to demonstrate their imagination and creativity, but also it encourages them to develop a deep learning approach in which they can acquire subject independent learning skills for learning, and through careful design of assessment, to practise and demonstrate their command of these skills. In order to encourage students to adopt a deep learning approach, the lecturer must lecture less, convince students of the intellectual aims of their course, and create opportunities, in lessons and outside, in which thinking can flourish.

Another positive for IR as a topic of study is that, if students can see the relevance of a course it generally aids their motivation. Most students taking IR courses are regular users of search engines and digital technologies in general, making the relevance of the material clear to their lives is thus not generally a problem for the instructor.

In terms of assessment it is relatively easy to set questions requiring students merely to memorize content delivered in lectures. They either learn a script delivered in lectures where students take down their own notes verbatim or use handouts made available either in printed form, or most likely made available for download from a module webpage. However, as we have seen this approach fails many of the generally desired outcomes of a university level education.

I have experimented with providing notes to students in advance of lectures enabling students to bring the handouts with them to the class so that they annotate them with personal additional comments, or making them available after the lecture to encourage students to listen and interact with the material during the lecture. Students, perhaps unsurprisingly, in my experience universally favour being provided with the materials in advance. While providing notes in this way seems to me to be pedagogically justified if the students make the intended use of them. It inevitably also creates problems, students who attend the class with the notes may not pay full attention, since they already have the notes, and students who are inclined to skip lectures can do so in the knowledge that the lecture is "on the web". Depending on the content of the notes, the correlation between the notes and the content of the lecture itself, the added value of actually attending or even participating in the lecture, and the assessment methods used in examinations or coursework, the students may feel fully justified in missing the class. There is after all little point in spending time in a class if you can just read it up from the notes in advance of the assessment and still be able to gain a distinction level mark. One could argue slightly dogmatically that students must attend classes, but this is academically and intellectually difficult to defend if there is visible evidence of the assessment outcomes showing there to be no reason for them to be there.

Furthermore, if continuous assessment assignments are based around reviews of existing approaches, e.g. writing review essays, many students take the approach of writing submissions that are highly derivative of recommended reading. Even if writing essays of this sort does not constitute plagiarism, and in some cases it probably does, it is not at all clear that students really gain much from such assignments. It seems to me much better to use these assessments to attempt to establish that the fundamentals of the subject have been understood, and then to explore the students' ability to break down a problem, and make use of their knowledge of the principles and techniques of the subject to address and report it in a creative way.

4. INFORMATION ACCESS

4.1. Background

While working at the University of Exeter, I developed a module called *Information Access*. This took IR as its hub, but introduced a range of related technologies for information searching within unstructured document collections and explored their integration to address information access tasks. The module was aimed primarily at final year undergraduate computer science students, and assumed a prerequisite of basic undergraduate statistics and introductory artificial intelligence methods. The general philosophy of teaching to these students within their programme of study was very much based on problem analysis and design and implementation of solutions. They were thus used to being challenged to learn a new subject by solving problems.

The technical focus of the module was to establish the unchanging issues and challenges of information access. In practice this was principally to convey the concepts of user information need, document collections, uncertainty of relevance and to make clear why accurate IR is difficult! The module introduced current techniques from IR very much from a practical rather than highly theoretical perspective to enable students to build effective prototype tools for IR (for example, well established ranking algorithms). Students completing the module were expected to able to follow a "recipe" for the construction of an effective small scale IR system, for example to follow easily the description appearing in [2].

Another aspect in design and delivery of module was to ensure the students understood that IR is a rapidly developing subject within which they can make novel and significant personal contributions. While this last aspect is perhaps obvious to experienced researchers, this is very much not the case for many undergraduate students. Students often do not realize that computing is a live subject within which they are free to propose, test and report new ideas. As part of the emphasis on the development of novel technologies the module also introduced evaluation for IR applications to enable testing of ideas and where appropriate comparison of potential solutions.

4.2. Syllabus

The module began by introducing standard IR topics including stopping, stemming, Boolean and ranked retrieval and term weighting. It then introduced the related information access topics of hypertext, information extraction, machine translation, speech recognition, information visualization, intelligent agents and summarization. This enabled exploration of a wide range integrated information access scenarios, including topics such as cross-language IR, spoken document retrieval (SDR), information exploration using graphical visualization, agent-based information discovery, and web searching. I should emphasize that each topic was covered at an introductory level of definition, establishing the fundamental challenges and problems (for example, speech recognitions produce errorful output, and will continue to do so for the foreseeable future, so what are the implications for SDR?), and in outline current methods used to implement each one.

4.3. Teaching Materials and Delivery

The wide range of topics covered meant that there was no suitable single set text that could be used. However, I sought to turn this into an opportunity for the students to develop information searching skills of their own. Key texts on each of the module topics were identified and made available in the university library. In addition key research papers were identified, some of these were tutorial style papers on the topics, while others represented examples of current research combining topics covered in the module. Copies of these papers were made available to the students as a module "reader" in the School library.

Students were also provided with a number of online handouts. These comprised an overview of the objectives and teaching approach of the module, and a detailed list of the papers and texts provided for the module. Notes from each lecture were also made available online. As discussed earlier there seems to be no ideal approach to the provision and timing of distribution of handouts. After some experimentation I found it generally to be most effective to make handouts available in advance of classes, but to keep them fairly brief and make clear that they represented a minimum requirement of knowledge on the topic, and that students were expected to read around the topics from the provided books and research papers. One or two of the provided papers were set as readings for each week and it was made clear that students would be expected to be familiar with these materials in lectures in following weeks. My reasoning here was that students at this stage of their study are generally not familiar with reading research materials, by setting specific readings they could start gaining familiarity with this style of writing, and the more extensive list of publications gave them a starting point for wider exploration, and also to help them identify the leading venues for publication of IR research. They were further encouraged to read beyond this list both by following references and searching the web as part of the module assessment.

The module was introduced to the class by explaining the general issues of searching unstructured information, starting with web as en example and moving on to integrated examples such as cross-language browsing of spoken content from document summaries. The assessment elements and methods were then explained, along with the expectations for reading of the module materials. For each topic the lectures introduced the principles of each topic, but as often as possible I encouraged students to think about the issues critically to recognize the inherent challenges in each topic for themselves. Creativity in solutions was encouraged by posing questions and working with the class to propose and explore solutions. For example, in order to achieve a particular IR task, such as cross language IR, what are the issues beyond monolingual IR that must be addressed?, how might these be addressed in a practical system? what are the strengthens, limitations, etc, of particular potential solutions?, and how might we test the effectiveness of the proposed methods?

4.4. Assessment

The module was assessed 75% by examination and 25% by a continuous assessment assignment. The examination was of a fairly traditional structure with free choice of 3 questions from 5 in two hours. Individual questions combined multiple topics from the module, this was because of the inherently integrated nature of the information access problems addressed, but also to restrict the opportunities for selective revision leading to high marks. The method of teaching and the general preparation of this class meant that it was possible to structure many of the examination questions around problems and scenarios. The beginning of each question required students to explain basic definitions or identify key issues; questions then moved to problem-based questioning where candidates needed to develop creative solutions and propose how these might be evaluated based on experience gained during the module.

The more novel element of the assessment was the continuous assessment assignment. Students were asked to select and consider one of a number of given "research" problems in information access. They were required to report their solution as a formally structured research paper. A template of a standard research paper format was explained: abstract, introduction, literature review, proposal, method of assessing proposal and anticipated possible results and conclusion, with properly formatted references. Submissions were required to be in this format with a prescribed maximum word limit. This form of assessment exercised a number of important skills, as well as testing understanding of IR. The abstract tested ability to write a succinct summary of a document. The introduction

needed to give suitable background and motivation, and detail of the topic and paper structure. The review required students to select relevant material and ignore non-relevant material from reviewed documents and then concisely express this material in an integrated fashion leading to the justification of a research proposal. A means of implementing and then evaluating the proposal must be described, and students must explore anticipated results and conclusions. The reference section must properly cite reviewed research papers, credit was given for wide reading of materials beyond that introduced in the lectures. A marking scheme making clear the requirements necessary of an ideal submission, and progressively weaker ones associated with each grade was included with the assignment. Students often found this assignment a very challenging exercise for a variety of reasons. Differing aspects challenged individual students, for example writing a review within a tight word limit, developing new ideas, or considering how to evaluate their ideas. Ultimately students generally agreed though that writing this assignment following the research paper template formed a very useful learning exercise. From an assessment perspective one particular strength of this assignment over a standard essay type review, is that students could not simply restate standard materials from books and papers.

5. FURTHER TEACHING OF INFORMATION RETRIEVAL

A number of students completing this module went on to undertaken final year projects under my supervision. The best of these were accepted for publication at international conferences [3][4][5], other students completed excellent projects which were not submitted for publication. I found that the module provided excellent preparation for these projects, in terms of basic subject knowledge, but also ability to conduct background research, creative thinking, and evaluation.

In addition, to the BSc module in Information Access, I later developed an MSc module in Natural Language Engineering. The first half was an introduction to natural language processing and the second half an introduction to IR. At MSc level I felt it important to adopt a more formal approach to the subject, and covered topics such as the derivation of the binary independence model and the Okapi BM25 model. The theory for this latter aspect was very challenging for the students, but it gave me the opportunity to explore the step from theory to practice for the model covered in [6]. In doing this I emphasized the importance of this type of theoretically motivated, but ultimately very pragmatic piece of work to IR and other areas of applied artificial intelligence and information engineering. This module also formed the basis of successful project work including [7].

REFERENCES

[1] Bligh, D. (1998) *What's the Use of Lectures?* (5th~edn). intellect.

[2] Robertson, S.E. and Sp\"{a}rck Jones, K. (1994, rev. 1996,1997,2006) *Simple, proven approaches to text retrieval.* University of Cambridge Computer Laboratory Technical Report no. 356.

[3] Jones, G.J.F., Quested, D.J., and Thomson, K.E. (2000) *Personalised Delivery of News Articles from Multiple Sources* In Proceedings of the Fourth European Conference on Digital Libraries, Lisbon, pages 340-343.

[4] Jones, G.J.F. and Gabb, S.M. (2002) *A Visualisation Tool for Topic Tracking Analysis and Development* In Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, pages 389-390.

[5] Jones, G.J.F. and Edens, R.J. (2002) *Automated Alignment and Annotation of Audio-Visual Presentations* In Proceedings of the Sixth European Conference on Research and Development for Digital Libraries, Rome, pages 276-291.

[6] Robertson, S.E. and Walker, S. (1994) *Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval* In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, pages 232-241.

[7] Edens, R.J., Gaylard, H.L., Jones, G.J.F., and Lam-Adesina, A.M. (2003) *An Investigation of Broad Coverage Automatic Pronoun Resolution for Information Retrieval* In Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), Toronto, pages 381-382.