Teaching of Web Information Retrieval: Web first or IR first?

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
Udine, Italy
mizzaro@dimi.uniud.it
http://www.dimi.uniud.it/mizzaro/

Abstract

When teaching Web Information retrieval (IR), a teacher has two alternatives: (i) to teach the classical pre-Web IR issues first and present the Web specific issues later; or (ii) to teach directly the Web IR discipline per se. The first approach has the advantages of building on prerequisite knowledge, of presenting the historical development of the discipline, and probably appears more natural to most lecturers, who have followed the historical development of the field. Conversely, the second approach has the advantage of concentrating on a more modern view of the field, and probably leads to a higher motivation in the students, since the more appealing Web issues are dealt with at course start.

I will discuss these issues, I will mention the approaches followed in the (rather few) Web IR books available, I will make some comparisons with the teaching of related disciplines, and I will also summarize my experience and some feedback from my students (I have been teaching a Web IR course for two Master's degrees in Computer Science and Information Technology at Udine University for the last two years; I had about twenty students each year; and I followed the first approach).

Keywords: Information Retrieval, Web information retrieval, literature survey, student feedback, student presentations.

1. INTRODUCTION: FROM IR TO WEB IR

Information Retrieval (IR) has been taught for decades in various curricula: computer science, computer engineering, information science, and library science. Even if research in the IR field was steadily progressing, the scenario was quite settled in terms of topics taught. Topics chosen in IR courses depended on the emphasis: more soft-science oriented curricula (e.g., library and information science) used to concentrate on the user side, studying cognitive aspects. More hard-science oriented curricula (e.g., computer science and engineering) concentrated on IR formal models, algorithms, data structures, and implementation techniques. In the last five-ten years, the situation has changed, the reason being Web and Web Search Engines coming. This change has brought to what might be considered a new discipline, namely Web IR. In Web IR the emphasis shifts on Web issues: search engines, crawling, models of the Web graph, link analysis, Web technologies, languages and protocols, search with mobile devices, etc.

Web IR is different from classical IR for two kinds of reasons: concepts and technologies. On the conceptual side, there are several Web specific issues which do not show up in classical IR circles, are not mere technicalities at all, and often rely upon well established fields: modeling of the Web graph (including graph theory, random graphs, scale-free and small world networks); links between PageRank/HITS and other notions of centrality in Social Network Analysis; crawling; low quality content and quality issues; heterogeneous users (not just librarians); etc. On the technological side, the lecturer that wants to follow the rapid evolution of search engines in the real world is overwhelmed with several technicalities. These might be disregarded as ephemeral and transitory, but I believe that this would be a double mistake. First, because the technical issues are often very effective in raising students' motivation; and second because often technical issues are very important and do make a difference. For instance, Google would not exists without the carefully designed architecture of the so called Google cluster (Barroso et al. 2003).

Given the above sketched scenario, the teacher of a course delivering both classical IR and Web IR issues has to choose between two alternatives: (i) to teach the classical pre-Web IR issues first and present the Web specific issues later, as an add-on; or (ii) to teach directly the Web IR discipline per se, dealing with classic IR issues when needed. The first approach has the advantages of building on prerequisite knowledge, presenting the historical development of the discipline, and, probably, appearing more natural to most teachers, who have followed the historical development of the field. The second approach has the advantage of concentrating on a more modern view of the field, and probably leads to a higher motivation in the students, since the more appealing Web issues are dealt with right at course start.

In this paper I discuss these issues. In the next section I mention the IR and Web IR books available, and I discuss the consequences on the two approaches. In Section 3 I make some comparisons with the teaching of related disciplines. In the following sections, to ground on a concrete case, I rely on the course I am teaching, summarizing my experience and some feedback from my students. Section 6 summarizes the paper and proposes some future developments.

2. BOOKS

In classical pre-Web IR courses, several reference texts were available; indeed, rather old texts like van Rijsbergen (1979), Salton and McGill (1984), Salton (1989), Blair (1990), Frakes and Baeza-Yates (1992), and Ingwersen (1992) were sometimes preferred to newer ones like Korfhage (1997), Marchionini (1997), Baeza-Yates and Neto (1999), Witten et al. (1999), Belew (2000), Grossman and Frieder (2004), van Rijsbergen (2004), and Ingwersen and Järvelin (2005). More soft-science oriented curricula usually adopted reference texts among Belew (2000), Blair (1990), Ingwersen (1992), Ingwersen and Järvelin (2005), and Marchionini (1997). More hard-science oriented curricula usually adopted texts among Baeza-Yates and Neto (1999), Belew (2000), Frakes and Baeza-Yates (1992), Grossman and Frieder (2004), Korfhage (1997), Salton (1989), Salton and McGill (1984), van Rijsbergen (1979), van Rijsbergen (2004), and Witten et al. (1999). Very few IR books deal with Web IR specific issues, and often in a very limited way.

To my knowledge, there are only two textbooks specifically devoted to Web IR issues: Chakrabarti (2003) and Levene (2006). Both of them start directly with Web IR issues, and IR topics are presented when needed. The former is more mathematical and conceptual oriented; the latter deals with technological issues to a greater extent. Both of them touch upon classic IR issues (the vector space model, the tf.idf weighting scheme, etc.), although these are not discussed and presented per se, but as means to search engines implementation.

Of course, the availability of a good textbook following either of the two approaches might lead to prefer one of the two. Also, it is not surprising that classical IR topics are much more covered than more recent Web IR topics: classic IR has a longer history and it is very difficult, if not impossible, to have a book which is fully up to date about the last technological issues.

Since there are several classical IR books, the IR first approach would benefit from Web IR books presenting the Web issues as an evolution of classic IR, and that take classic IR techniques and methods as prerequisite. Perhaps surprisingly, no such books are available: either Web IR issues are dealt with marginally in IR books, or Web IR issues are introduced right from the beginning of a Web IR book. This seems to be a serious hindrance to the IR first approach.

However, there is a solution to this problem: all “classic” Web IR papers, like, for instance, Albert and Barabasi (1999), Albert et al. (1999), Barroso et al. (2003), Bharat et al. (2001), Brin and Page (1998), Broder et al. (2000), Cho and Garcia-Molina (2000), Fetterly et al. (2004), Gulli and Signorini (2005), Kleinberg (1999), Lawrence and Giles (1998), Lawrence and Giles (1999), and Page et al. (1998), are freely available on the Web, and are good substitutes for Web IR books. As a matter of fact, since Web IR is in its beginning, these papers are usually free from too technical and difficult details and methods, and they can be easily read and followed by students. Reading Web IR research papers has also the positive effect of making students more aware of the scientific method, of how researchers work, and of the main conferences and journals of the field. Finally, the most recent technological developments will not be found in books: accessing resources on the Web (like, for instance, major search engines pages, or specialized Web sites like, for instance, searchenginewatch.com or searchengineshowdown.com) is necessary. Thus, the IR first approach does not suffer from the lackness of Web IR books that take IR as a prerequisite.

Turning to the Web IR first approach, we see that it is supported by two books. Of course, two books are not many, but in my opinion they are quite good books, although in different respects. Moreover, the situation will change for sure in the near future, with many more Web IR books being published.

Summarizing this short analysis, both approaches are compatible with currently published books.

3. RELATED DISCIPLINES

Another way to study the alternative between IR first and Web IR first is to analyze what is done in other fields. I see two approaches here. Well established disciplines (calculus, algebra, physics, chemistry, etc.) are more detached from the historical development. More recent disciplines (typical examples can be found in computer science subtopics, like programming languages, operating systems, database theory, etc.) seem to be somehow reminiscent of the historical development of the field. For instance, when teaching programming language courses it is not uncommon to follow the historical evolution of the languages: Assembly, Fortran and Cobol, Algol and other structured programming languages, Object oriented, etc.

Another similar and well discussed issue concerns introduction to programming courses. Among several alternatives (IEEE & ACM 2001, pp. 29-30), two are of interest here: whether to adopt an imperative-first approach and teach structured and procedural programming first, moving later to object oriented issues perhaps going through the Abstract Data Type concept, or whether to choose an object-first approach, and teach programming directly in object oriented terms.

Both these approaches are feasible and supported by several textbooks, and none of them is criticized in IEEE & ACM (2001). During the last 10 years or so I have been adopting the imperative-first approach in first year introduction to programming courses for Engineering, Computer Science, and Information Technology students, and I am satisfied with this approach.

4. MY EXPERIENCE

I will now turn to my (somewhat limited) teaching experience relevant to the Web IR field. I have been teaching a Web Information Retrieval course for two Master's degrees in Computer Science and Information Technology at Udine University for the last two years. The duration of the course was one term (48 class hours). I had about twenty students each year; 17 last year plus 17 this year have already passed their exam; about 5 more each year (10 more in total) have attended the lectures but have not passed the exam yet. The majority of the students were graduate, but about 5 each year were undergraduate. I followed the IR first approach.

4.1. Syllabus

My Web IR course was divided into two parts; the syllabus is (there have been only minimal changes between the two years):

- IR
  - Introduction to IR.
  - IR models: boolean, vector space, probabilistic, latent semantic indexing, neural nets.
  - Inverted index: structure, construction, use.
  - Query languages.
  - Reformulation techniques.
  - Human Computer Interaction and IR: examples of user interfaces for IR.
  - Clustering: main algorithms, usage in IR.
  - Evaluation: test-collections and user studies.
- Web IR
  - Statistics on Web and Web users.
  - Models of the Web Graph (random graphs, scale free and small world network, power laws, bow-tie, Web host graph).
  - Link analysis for ranking (PageRank, HITS)
  - Spam, duplications, mirrors.
  - Search engine architecture.
  - Crawling.

4.2. Students presentations

Besides normal lessons, the course also featured some students presentations as simple term projects, on voluntarily basis. These presentations lasted one hour (or less) and were either on specific conceptual issues (i.e., read one or more research papers and summarize it/them) or on technological issues (i.e., analyze a specific system/technology and summarize it). So far, students presentations have been on the following topics: link analysis for ranking algorithms; stemming algorithms; Google desktop; Spotlight; IR in P2P networks; Search Engine Optimizers (SEO), improving a Web site to obtain a higher visibility to search engines; video retrieval concepts and systems (Google Video, YouTube, Yahoo!); Google architecture (massive parallelization, hardware management, power consumption and refrigeration, etc.); Self Organizing Maps; Vivisimo; and Average Distance Measure (ADM), a specific IR effectiveness measure.

I have found students presentations extremely positive. When focused on conceptual topics, they were useful to the speaker and understandable to the audience; when concerning technological issues they were sometimes even more detailed than what I could have done; the audience looked very interested; and during the exams, questions on the main concepts were correctly answered.

5. STUDENTS FEEDBACK

Some feedback has been gathered from last two years students by means of an anonymous questionnaire, administered either on paper during a lecture or sent by email. The questionnaire consists of 21 questions, divided into two parts. The first part (6 questions) aims at collecting students' background knowledge before their attended the lectures; the second part gathers students' opinion about the course topics included (or not included) in the course, difficulty of specific topics, opinions on students presentations, etc. Most of the answers are on a five points Likert scale; some of them are in free text; and there are two semantic differentials collecting students' impressions on the IR and Web IR parts of the course. The main findings are:

- Twenty-three students returned the questionnaire (52% of the 44 potential respondents), 9 of them from the 04/05 academic year, and 14 of them from the 05/06 year.
- Before course start, students knew what a search engine is (average of 3.9 on the 1–5 five points scale, i.e., about one unit above the medium value, which is 3). The concepts of crawler (2.7) and inverted index (1.9) were less known.
- Three questions were aimed at understanding if classical IR is somehow less enjoyable than Web IR, which turned out not to be the case: Web IR is just slightly more interesting than IR (3.1); Web IR issues are not perceived as more difficult than IR (2.8); and students desire of more emphasis on Web IR topics is very small (3.4).

  Further evidence is given by the results of the two semantic differentials, that ask if the IR (Web IR) part of the course is interesting/boring, difficult/easy, significant/insignificant, unpleasant/pleasant, simple/complex, useless/useful. The average results (on a seven points scale and "normalized" on the positively polarized items - higher values mean higher interest, ease, etc.) are summarized in Table 1: IR is as easy, significant, and simple as Web IR, and Web IR is judged more interesting, more pleasant, and slightly more useful than IR.
- Coming to the main question dealt with in this paper, two questionnaire answers support the "IR first stance": students do perceive the usefulness of classic IR to understand Web IR (4.3 - the higher value among all questions); also, this value is even higher than IR usefulness to understand historical development (4.2), which is probably unquestionable.
- Students talks are considered rather useful (3.8). Other questionnaire items confirm that the syllabus is generally perceived quite positively: more Web issues are requested only to a limited amount (3.4); deepening of specific topics is only slightly suggested (3.4); and less emphasis on some issues is requested to an amount lower than the median value (2.9).

	INTERESTING	EASY	SIGNIFICANT	PLEASANT	SIMPLE	USEFUL
IR	4.7	3.5	5.3	4.9	2.4	6.1
WEB IR	5.7	3.5	5.2	6.1	2.6	6.6

TABLE 1: ANSWERS ON IR AND ON WEB IR.

6. CONCLUSIONS AND FUTURE WORK

In this paper I have mainly discussed whether a Web IR course should deliver Web IR issues right at course start (Web IR first approach) or after a classical-historical description of classic IR issues (IR first approach). I have not found yet any hard evidence suggesting that the IR first approach is wrong; conversely I do have several clues that the IR first approach is effective: (i) IR issues are a strong basis for Web IR issues; (ii) even if there are no books fully covering the IR-first approach, the research papers are accessible and understandable to students; using research papers has also the positive side effect of making the students acquainted with the scientific literature of the field; (iii) an historical-like approach (as IR first is) is followed in several other disciplines; and (iv) accordingly to the results of a small survey, students do not feel unmotivated by having to wait the second half of the course for Web topics.

Therefore I will use the IR first approach again next year. Anyway, I understand that for my personal teaching "style" this is the most natural approach (I am teaching introduction to programming in the same way, and during my university studies I was taught several computer science topics in the same way), which might not be true for all teachers. I will be happy - and I will welcome discussion - to collect feedback from other lecturers during the workshop.

In the above discussion, I also had touched upon the topics I am currently teaching in my course, briefly describing the syllabus of the course and providing some evidence that students agree with it. I also briefly described the IR and Web IR books available, and maintained the effectiveness of having class presentations from students on specific technical and/or conceptual issues.

To mention at least a negative side of the IR first approach, I remark that it does not fit well with the students presentations: since most of the talks are on Web IR issues, these happen to be chosen quite late, towards the end of the term, and there is some organizational problem in finding an appropriate schedule before course end.

To close this paper, I suggest considering (besides - or in place of - a journal special issue) a book on Web IR. The book could be a Wikibook (http://wikibooks.org), perhaps within the Wikiversity project (http://en.wikibooks.org/wiki/Wikiversity), written collaboratively among all the teachers - and students - of (Web) IR courses and freely available on the Web. The book could also contain a collection of specific term projects and, in my opinion, should be suited to the IR first approach.

REFERENCES

R. Albert and A-L. Barabasi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. http://arxiv.org/format/cond-mat/9910332.

R. Albert, H. Jeong, and A-L. Barabasi. Internet: Diameter of the world-wide web. *Nature*, 401: 130–131, 1999. http://www.nature.com/nature/journal/v401/n6749/abs/401130a0_fs.html.

R. Baeza-Yates and R. Neto. *Modern Information Retrieval*. ACM Press, 1999.

L. A. Barroso, J. Dean, and U Hölzle. Web search for a planet: The Google cluster architecture. *IEEE Micro*, 23(2):22–28, 2003.

R.K. Belew. *Finding Out About*. Cambridge Univ. Press, 2000.

K. Bharat, B-W. Chang, M. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. *In IEEE International Conference on Data Mining (ICDM '01),* 2001.

D. C. Blair. Language and Representation in Information Retrieval. Elsevier, 1990.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7*, pages 107–117, 1998. http://www-db.stanford.edu/˜backrub/google.html.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000. http://www.people.cornell.edu/pages/dc288/Paper1.pdf.

S. Chakrabarti. *Mining the Web*. Morgan Kaufmann, 2003.

J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In *Proceedings 26th VLDB*, 2000. http://rose.cs.ucla.edu/˜cho/papers/cho-evol.pdf.

D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. *Software Practice & Experience*, 34:213–237, 2004. http://research.microsoft.com/research/sv/sv-pubs/pageturner-spe2004.pdf.

W.B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice- Hall, 1992.

D. A. Grossman and O. Frieder*. Information Retrieval: Algorithms and Heuristics*. Springer, 2nd edition, 2004.

A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW14*, 2005. http://www.cs.uiowa.edu/˜asignori/web-size/.

IEEE & ACM. The Joint Task Force on Computing Curricula — Computing Curricula 2001: Computer Science — Final Report, December 15 2001. http://acm.org/education/curric_vols/cc2001.pdf.

P. Ingwersen. Information Retrieval Interaction. Taylor Graham, 1992.

P. Ingwersen and K. Järvelin. *The TURN: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632, 1999. http://www-courses.cs.uiuc.edu/˜cs591han/papers/klei99.pdf.

R. R. Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, 1997.

S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280:98–100, 1998. http://www.sciencemag.org/cgi/content/full/280/5360/98.

S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.

M. Levene. *An Introduction to Search Engines and Web Navigation*. Addison Wesley, 2006.

G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1997.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. http://www-db.stanford.edu/˜backrub/pageranksub.ps, 1998.

G. Salton. *Automatic Text Processing — The Transformation, Analysis, and Retrieval of Information by Computer*. Addison Wesley, 1989.

G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1984.

C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes — Compressing and Indexing Documents and Images*. Morgan Kaufmann, 2nd edition, 1999.