DD2475 Information Retrieval

Course Analysis, ir10

Hedvig Kjellström and Johan Boye

This analysis has been performed by Hedvig Kjellström and Johan Boye and is based on an online student questionnare, and on discussions with students during the course.

Course Data

The course was given for the first time during periods 2, 2010 and 3, 2011, with course leader Hedvig Kjellström. 9 students took the course. The teaching activities consisted of

- 14 lectures of which 7 were given by Hedvig Kjellström, four by Johan Boye, one by Viggo Kann, and two by guest lecturers
- 14 two-hour computer hall sessions held by Hedvig Kjellström and Johan Boye , where students could ask for help with the programming tasks

The theoretic part of the course (3 hp) was examined with a written exam in the end of period 2. The students also performed three computer assignments (3 hp) which were examined orally in periods 2 and 3, and one larger project (3 hp) which was examined by a interactive poster session and a written report in the end of period 3.

The book used in the course was C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008. We also used the introduction of U. von Luxburg, A tutorial on spectral clustering, *Statistics and Computing* 17(4), 2007.

The book was (rightfully we think!) highly appreciated by the students. One student wrote in the questionnare: *" A very good reference book."*

Course Name

This course covers the new and expanding area of Information Retrieval, i.e., the science of finding information in large quantities of unordered data, such as on the Internet.

The Swedish course name, Informationssökning, is very unfortunate. Although the established Swedish translation to Information Retrieval, it is still unknown to the students due to the novelty of the field. To the public, this term rather refers to manual information search – there is even a course with this name at KTHB! Consequently, a number of librarians applied for the Information Retrieval course in 2010.

Swedish students are repelled from the course because they misunderstand the Swedish name. One student wrote in the questionnare: *"Informationssökning är alldeles för likt den andra 'Programsammanhållande' kursen på 1.5 HP. Den här kursen har svårt att få ett*

bra varumärke p.g.a. detta. Ett annat namn behövs, som verkligen säger att det är insidan på sökmotorn som behandlas. Inte 'Sök på internet, 1.5 HP'. Varumärket för en kurs är extremt viktigt för hur populär den blir. Elever väljer kurser som låter avancerade, mest utifrån namnet."

Action points:

We suggest that the course changes name to

Search Engines and Information Retrieval Systems
Sökmotorer och informationssökningssystem

to emphasize that it treats algorithms for automatic retrieval, not manual search.

Learning Outcomes

The course is based on Christopher D. Manning's book and course at Stanford. We started off with a course that covered the whole book, and consequently formulated the following learning outcomes and course content:

After completing the course you will be able to:

- Explain the concepts of indexing, vocabulary, normalization and dictionary in Information Retrieval
- Define a boolean model and a vector space model, and explain the differences between them
- Explain the differences between classification and clustering
- Discuss the differences between different classification and clustering methods
- Choose a suitable classification or clustering method depending on the problem constraints at hand
- Implement classification in a boolean model and a vector space model
- Implement a basic clustering method
- Give account of a basic spectral method
- Evaluate information retrieval algorithms, and give an account of the difficulties of evaluation
- Explain the basics of XML and Web search

Content
Basic and advanced techniques for information systems: information extraction; efficient text indexing; indexing of non-text data; Boolean and vector space retrieval models; evaluation and interface issues; XML, structure of Web search engines; clustering, classification; spectral methods, random indexing; data mining.

We found during the course that XML was somewhat to the side of the main course subject, and believe that the students can acquire XML programming competence autonomously once they have a grasp of the theory of Information Retrieval.

We also found that the course put a too heavy focus on classification and clustering; these subjects are treated in DD2431 Machine Learning , taken by the great majority of the students in the Information Retrieval course.

Students also found the course scope to be a bit to wide. One student wrote: *"In my opinion I think the whole book of Manning need not be covered. Taking into account that some of the topics are discussed in other courses as well such topics can be removed from the curriculum."*

Another student commented: *"Some of the labs (especially the last) were not really 'Information Retrieval', just basic algorithms. Computer scientists that choose the course have already done this and will learn nothing."*

Action points:

We suggest to decrease the focus on clustering and classification, and to remove XML. New learning outcomes and content could be:

After completing the course you will be able to:

- **Explain the concepts of indexing, vocabulary, normalization and dictionary in Information Retrieval**
- **Give an account of different text similarity measures, and select a similarity measure suitable for the problem at hand**
- **Define a boolean model and a vector space model, and explain the differences between them**
- **Implement a method for ranked retrieval of a very large number of documents with hyperlinks between them**
- **Evaluate information retrieval algorithms, and give an account of the difficulties of evaluation**
- **Give an account of the structure of a Web search engine**

Content
Basic and advanced techniques for information systems: information extraction; efficient text indexing; indexing of non-text data; Boolean and vector space retrieval models; evaluation and interface issues; structure of Web search engines.

Below we describe how the learning activities and examination is changed to reflect this.

Course Activities

The course activities consisted of 14 lectures and 14 computer hall sessions, once a week during both periods. Since the course followed Manning's book closely, the lecture series covered all chapters in the same order as they appeared in the book:

Lecture 1: Given by Hedvig Kjellström. Manning Chapter 1
Lecture 2: Given by Johan Boye. Manning Chapter 2, 3
Lecture 3: Given by Johan Boye. Manning Chapter 4, 5
Lecture 4: Given by Hedvig Kjellström. Manning Chapter 6, 7
Lecture 5: Given by Hedvig Kjellström. Manning Chapter 8, 9
Lecture 6: Given by Hedvig Kjellström. Manning Chapter 10
Lecture 7: Given by Hedvig Kjellström. Manning Chapter 11, 12
Lecture 8: Given by Johan Boye. Manning Chapter 13
Lecture 9: Given by Hedvig Kjellström. Manning Chapter 14, 15.3
Lecture 10: Given by Hedvig Kjellström. Manning Chapter 16.1-4, 17.1-3
Lecture 11: Given by Viggo Kann. Manning Chapter 18, Luxburg Sections 1-4
Lecture 12: Given by Johan Boye. Manning Chapter 19, 20, 21
Lecture 13: Guest lecture by Jussi Karlgren, SICS and Hercules Dalianis, SU DSV (the last talk cancelled due to illness)
Lecture 14: Guest lecture by Simon Stenström, Findwise and Magnus Rosell, Recorded Future

None of the lecturers are specialized in Information Retrieval; the research area of Hedvig Kjellström is Computer Vision and Machine Learning, and Johan Boye specializes

in Language Technology and Linguistics. The lectures leaning towards language processing were therefore given by Johan Boye, and the lectures with learning content were given by Hedvig Kjellström. Both lecturers were present at all lectures, to make sure that the lecture series was held together despite the wide scope of the course.

The students appreciated the lecturing style with plenty of interaction: *"The interactive way of teaching (small exercises during the lecture) was very effective as we were able to relate to the topic discussed. Definitely retain the same lecturers."*

The differences in lecturer focus were also appreciated: *"The combination of two lecturers were great, with regard to their expertise domain and their discussion during the course."*

The guest lecturers from industry were a success, and will definitely be retained in the coming years. One student liked the *"different kinds of lectures, like in the end we have some people from companies speaking"*.

One student asked us to more clearly relate to the book in the lectures: *" Concerning the slides, numbering, putting title of the cahpter in each slide could be useful for the student to know in each moment what is being said."* We will look into that.

As discussed above, we found that XML (Chapter 10) should be removed from the course curriculum, and that there should be less focus on classification and clustering as this is covered in earlier courses. In return we suggest to put more focus on retrieval of documents with hyperlinks – i.e., web pages. This means that there is a clearer focus on web search, something that is reflected in the new suggested course name.

We noted that there were far to much computer hall time scheduled. Depending on the number of students next year, the time for programming help could easily be lowered to 50% or even 20%.

Action points:

We suggest to decrease the number of scheduled computer hall time significantly., and compensate by providing individual help to the students upon request.

We suggest to change the lecture plan according to the following:

Lecture 1: Given by Hedvig Kjellström. Manning Chapter 1
Lecture 2: Given by Johan Boye. Manning Chapter 2, 3
Lecture 3: Given by Johan Boye. Manning Chapter 4, 5
Lecture 4: Given by Hedvig Kjellström. Manning Chapter 6, 7
Lecture 5: Given by Johan Boye. Manning Chapter 21
Lecture 6: Given by Hedvig Kjellström. Manning Chapter 8, 9
Lecture 7: Given by Hedvig Kjellström. Manning Chapter 11, 12
Lecture 8: Given by Hedvig Kjellström. Manning Chapter 14-17 (parts of)
Lecture 9: Given by Johan Boye. Manning Chapter 19, 20
Lecture 10: Guest lecture given by Viggo Kann
Lecture 11: Two guest lectures given by external people
Lecture 12: Two guest lectures given by external people

There are fewer lectures in the new plan. The general idea is that Lectures 1-7 cover the basic theory of the course and prepare for the written exam, while Lectures 8-12 provide outlooks that are valuable for the projects.

Specifically, the lecture about XML (old Lecture 6) has been replaced with a lecture about distance measures for documents with hyperlinks (new Lecture 5). The lectures about clustering and classification (old Lectures 8-11) have been compressed to one (new Lecture 8). Viggo Kann's lecture (old Lecture 11) is suggested to turn into a guest lecture (new Lecture 10) with outlooks on Language technology. The two lectures with external speakers are kept, preferrably with the same people.

Examination

The written exam (3 hp, graded A-F) was a success we think, and will be retained in the same format and with the same content: a theory part (aids: none) with 5 questions giving 20 points altogether, and a problem part (aids: all written material) with 5 questions giving 30 points altogether. Grading was done based on the total number of points. The scope of the exam was the content of the first 7 lectures.

In the exam December 13, the results were: 2 P (PhD students), 1 A, 3 B, 1 D, 1 Fx and 1 F. The Fx was completed orally to an E. An oral exam was given March 17, with the result: 1 B. As of March 17, the results are thus: 2 P, 1 A, 4 B, 1 D and 1 E. All 9 students have thus passed the exam.

There were three computer assignments (3 hp, graded P/F), which were examined orally in front of the computer during scheduled computer hall hours. The students worked in pairs or alone, and were allowed to select their partner themselves.

The assignments were arranged as:

Assignment 1.1: Basic Inverted Index
Assignment 1.2: Positional Indexing
Assignment 1.3: Large Inverted Indexes: Storing, Retrieving and Merging

Assignment 2.1: Index Compression
Assignment 2.2: Ranked Retrieval

Assignment 3.1: KNN Classification
Assignment 3.2: K-means Clustering

The students were generally happy with the computer assignments. One student wrote: *"The lab sessions with provided template codes were extremely useful, letting us to relieve from excessive coding."* Another commented: *"The code skeletons were really good since the students didn't have to worry about all the really tedious overhead work with IO and such. Other courses often fail here, making even simple labs take way too much time. Keep that up!"*

For the coming year, three changes will be made. Firstly, the code skeleton for the lab course will be (even) better prepared, with a user interface common to all assignments, and the possibility to add all modules to the same code, instead of making copies of the code for each new assignment. A new corpus with interlinked documents will also be collected, e.g., a part of Wikipedia.

Secondly, the third assignment will be replaced with a problem on retrieval of linked documents, to reflect the changes in learning outcomes (above). This is also in line with

what students suggest in the questionnare: *"I think more materials can be put in Lab sessions, specially those related to last lectures."* and *"I was expecting to implement more heuristics. There could possibly be another extension to the tf-idf lab."*

Last, one student suggests firmer deadlines for computer assignments: *"For the assignments, a clear deadline date could be helpful, for forcing oneself to finish this assignment in a limited time."* We think this is a good idea and will implement that. One sign that this is necessary is the results. As of now, the results of the lab course is 3 P (out of 9). Of the students not finished, 3 are missing one assignment, 1 is missing two assignments, and 1 is missing three assignments.

The students were also required to perform a project (3 hp, graded A-F), examined by an interactive poster session and a written report. Grading was done according to the following: Five compulsory and four additional criteria on the poster presentation and report were defined – for more information visit the course homepage at www.csc.kth.se/DD2475. To pass, all compulsory criteria have to be fulfilled. Higher grades are reached by fulfillment of additional criteria.

The projects were defined by two of the guest lecturers. The project formulations were intentionally very brief, as specification of the project was part of the task.

One thing to attend to is continuous supervision of the project work. Students commented: *"Great projects. Perhaps a little too loose, making it too easy. It's great to have contact with the industry." and "Concerning the project, some guidelines in the middle of the February will be nice so that the student know is going in the right way."* This indicates that controls, e.g., in the form of project diaries or progress reports could be useful to ensure project quality and to stimulate the students to work continously.

As of March 11, the results of the project are: 5 A and 2 C. Two PhD students are working on a slightly longer report, which is due in period 4, 2011.

The grade of the course is assigned as the mean (rounded up) of the project and written exam grades. Four students have finished all parts of the course, with the following grades: 2 A, 2 C.

Action points:

We suggest to replace Assignment 3 with an assignment about retrieval of documents with hyperlinks. The assignmens will be prepared with a more elaborate code skeleton. We also propose to move the examination of all assignments to the first period, to allow for more focus on the project.

The assignments will be arranged in two groups as:

Assignment 1.1: Basic Inverted Index
Assignment 1.2: Positional Indexing
Assignment 1.3: Large Inverted Indexes: Storing, Retrieving and Merging
Assignment 1.4: Index Compression

Assignment 2.1: Ranked Retrieval with tf-idf
Assignment 2.2: Ranked Retrieval with PageRank
Assignment 2.3: Ranked Retrieval with tf-idf and PageRank

Assignment 1 covers Lectures 1-3 and can be examined one week after Lecture 3. Assignment 2 covers Lectures 4-5 and can be examined one week after Lecture 6 or the week after.

The project will run during during the second period as before, but we will more closely monitor the work progress, using e.g., project diaries or progress reports.

The first period will then cover 6 hp (assignments, written exam), and the second period will cover 3 hp (project).