Teaching Information Retrieval With Web-based Interactive Visualization

Peter Brusilovsky
School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260.
E-mail: peterb@pitt.edu
Jae-wook Ahn
School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260.
E-mail: jaa38@pitt.edu
Edie Rasmussen
School of Library, Archival and Information Studies, University of British Columbia,
Vancouver BC V6T 1Z1, Canada. E-mail: edie.rasmussen@ubc.ca

Interactive visualization is a powerful educational tool, which has been used to enhance the teaching of various subjects from computer science to chemistry to engineering. This paper describes the use of interactive visualization tools in the context of a graduate course in information retrieval, to demonstrate two well-known retrieval models, the Boolean model and the vector space model. The results of five classroom studies with these tools are reported. The impact of the tools on student learning, as well as student attitudes toward the tools, were investigated. The results of the classroom studies indicate that use of interactive visualization in a homework context can result in significant growth of knowledge. The majority of the students recognize the value of interactive visualization and recommend its use in the context of information retrieval courses. The study also demonstrated that visualization focusing on less known and harder to understand topics causes a larger growth of knowledge and is perceived as more useful. This result suggests placing higher priority on the development of visualization tools for harder to understand topics.

Keywords: information retrieval, education, visualization, evaluation, user study

Introduction

Interactive visualization is a powerful educational tool. Visualization can provide a clear visual metaphor for understanding complicated concepts and uncovering the dynamics of important processes that are usually hidden from the student's eye (Gordin & Pea, 1995). Visualization has been used to enhance the teaching of various subjects ranging from chemistry (Evans, Yaron, & Leinhardt, 2008) to biology (McClean, et al., 2005) to physics (Perkins, et al., 2006). Computer science is one of the most active application areas for educational visualization research (Naps, et al., 2003). In computer and information science (CIS) education, visualization is used almost exclusively in programming and data structure courses. We can name dozens of papers devoted to visualization of program execution on several levels from machine-level languages (Butler & Brockman, 2001) to high-level languages (Domingue & Mulholland, 1998; Haajanen, et al., 1997; Levy, Ben-Ari, & Uronen, 2003; Tung, 1998) to algorithms and data structures (Hundhausen, Douglas, & Stasko, 2002; Rößling, Schüer, & Freisleben, 2000). Our claim is that In-

ISSN: 0748-5786

formation Science courses could benefit from this powerful technology.

This paper presents our research on using interactive visualization in the non-traditional context of information retrieval courses. Information retrieval has been in the curriculum of many computer, information, and library science departments for more than 30 years. With the maturity of the World Wide Web, information retrieval became an important practical subject. Elements of information retrieval are now taught to students of many different specialties. We think that information retrieval provides an interesting and important application area for exploring the power of interactive visualization. Over the last six years we developed and made publicly available a range of Web-based interactive visualization tools supporting various concepts taught in information retrieval courses. These tools were used for several years by faculty members at the University of Pittsburgh to teach both graduate and undergraduate information retrieval courses. To evaluate the impact of our visualization tools and to solicit student feedback, we ran several classroom studies. The developed set of tools is available on the project home page (http://ir.exp.sis.pitt.edu/res2/resources.php). The tools, which are running on our servers, could be used by anyone interested in teaching or learning information retrieval. This paper summarizes the results of our multi-year effort. To demonstrate our ideas, we present two examples of these interactive Web-based visualization tools for information retrieval. We also report the results of classroom studies evaluating them. To conclude, we discuss the results and prospects for using interactive visualization in the context of information retrieval courses.

Interactive Visualization for Information Retrieval

One of the secrets of the power of interactive visualization is its ability to uncover and present in detail processes that are typically hidden from students' eyes. Being interactive, visualization programs allow the students to explore these processes step-by-step, with different parameters, gaining as a result a deep understanding of the processes and the concepts behind them. Every field has its critical processes and concepts, which could be better understood with the use of visualization.

The core of a traditional information retrieval (IR) course is a set of models, algorithms and technologies for processing, storing and retrieving textual information. Traditional presentation of this core usually starts with several IR models (such as the Boolean, vector, and probabilistic models and several variations of them) and then follows by explaining how the information is organized and retrieved in each of these models (Baeza-Yates & Ribeiro-Neto, 1999; Korfhage, 1997). The process by which information is retrieved in different models is one of the hardest topics in an IR course for students to grasp, despite being formalized and well understood by the IR research community. We have observed that even Boolean information retrieval, the simplest of the models, is difficult for many students. At the same time, traditional educational tools—research or commercial IR systems—offer little educational help. The process of retrieving information has several steps, from entering the query to matching the query to the documents to prioritizing the results. In an IR system (even an educationally-oriented one) all these steps are hidden from a user: the only thing that a user can observe is the final results—a list of ordered documents. This is exactly a context that could benefit greatly from the use of interactive visualization and this realization formed the starting point for our research (Brusilovsky, 2002). Over the first years of our project we developed and explored interactive visualization tools to visualize the process of retrieving information in several known models: Boolean, fuzzy, vector, and extended Boolean (see Baeza-Yates & Ribeiro-Neto, 1999 and Korfhage, 1997 for the description of these classic models). Since these visualization programs were the first to be developed, we had more opportunities to improve them over the years and to evaluate them in the classroom. For the purpose of this paper we chose two of these model visualization tools to demonstrate the ideas of interactive visualization in the context of IR courses. The following subsections present the most recent versions of interactive visualization programs for the Boolean and vector IR models. The next section reports the results of classroom studies with these models.

Interactive Visualization of the Boolean Information Retrieval Model

The Boolean IR model is the oldest and the simplest of the IR models. In this model, a query is formed by a set of elementary queries (usually *keywords*) connected by Boolean operators such as AND, OR, and NOT. The mechanism of this model is set theoretical. Every query is associated with a set of matching documents. For an elementary query such as a keyword, the set of matching documents is simply all documents indexed by this keyword. To obtain the set of *matching documents* for two queries connected by a Boolean operator, one has simply to perform the corresponding set operation on their matching sets (i.e., set intersection for AND, complement for NOT, etc.). Thus in several steps, a matching set for any complex Boolean query can be found.

While it all sounds quite simple and clear, we have found that many of our students have problems understanding how Boolean matching works. Our talks with students have indicated that one of the sources of their troubles is the failure to perceive Boolean operators as operations on sets of matching documents. This is a known problem for everyday users who confuse the informal use of AND, OR, and NOT in everyday language with their formal meaning in set theory and Boolean logic. Surprisingly, we also discovered that some students with good programming backgrounds (i.e., those who have routinely used Boolean operators for writing conditional expressions in their programs) still have problems transferring their knowledge of these operators to the set theory context.

In developing an interactive visualization environment for the Boolean IR model we were trying to achieve two goals: to provide a helpful visual metaphor and to visualize the process of Boolean matching step by step. Figure 1 presents an interface for our environment. The core of this interface is a set of all documents visualized in a table (one document per row). For a sample document set in our system we choose textbook surrogates since this is the type of document most familiar to information science students. The goal of this visual representation is to help students to understand the core principle of this model—every query is associated with a particular *subset* of all documents. Showing the set of all documents on the screen makes it easy to demonstrate different subsets of the whole set as sets of differently colored rows of the table.

The students explore Boolean matching by writing and executing simple Boolean queries: pairs of elementary queries (terms) connected by a single Boolean operator (OR, AND), or a pair of operators (AND NOT), which simulate set difference. The results of an executed query are shown through row and cell coloring in the document table. The final set of documents is visible as a set of green rows (documents 2, 8, 12 in Figure 1). This is what a standard Boolean search engine would return. The visualization tool, however, attempts to show

Term 1		Condition	Term 2		
Any field	= Java	AND	Any field	=	1999

Id	Author	Title	Year	Publisher	Term 1	Term 2	Result
1	Brusilovsky	Adaptive Hypertext and Hypermedia	1998	Addison-Wesley	false	false	false
2	Deitel	Java. How to program	1999	Java soft	true	true	true
3	Sedgewick	Algorithms in C	2000	Addison-Wesley	false	false	false
4	Deitel	Java. How to program	1995	Java soft	true	false	false
5	Brusilovsky	Adaptive Hypermedia and Adaptive Web-Based Systems	2000	Addison-Wesley	false	false	false
6	Kendall	Systems Analysis and Design	2000	Oxford univ	false	false	false
7	Alter	Information Systems	1999	Oxford univ	false	true	false
8	Ladd	Using HTML 4, XML, and Java 1.2	1999	Java soft	true	true	true
9	Shelly	Using Computers as a Gateway to Information	1995	O'Reilly	false	false	false
10	Flanagan	Java in a Nutshell	1996	Java soft	true	false	false
11	Flanagan	Java in a Nutshell	1997	Java soft	true	false	false
12	Flanagan	Java in a Nutshell	1999	Java soft	true	true	true
13	Korfhage	Information Storage and Retrieval	1997	O'Reilly	false	false	false

Figure 1. Boolean Model Environment. Visualization of matching for a simple Boolean AND query. Documents matching the first elementary query are highlighted.

more by decomposing the standard "black box" matching process into two parts: (1) the process of matching an elementary query to the set of the documents and (2) the process of obtaining a new set from contributing sets by applying different Boolean operators. To achieve this goal, the tool extends the document table with three colored columns of cells. The first and the second column show the results of matching each of the documents to the first and the second query terms correspondingly. Documents matched to each term are marked as true in the corresponding column. To make the set of matched documents more visible, documents matched to the first term are marked by a red cell background and documents matched to the second term are marked with a blue background. The third column shows the results of matching the whole query using truth values and a green color for matching documents. The color-coding makes the two steps of the Boolean query matching process more transparent. For example, the student can see that only documents 2, 8, and 12, which matched to both elementary terms, are included in the final set. In contrast, documents 4, 10, and 11, which matched only to the first term and document 7, which matched only to the second term, are not included (see Figure 1).

Beyond the term-based search shown on Figure 1, the Boolean IR visualization tool has several other functionalities. In particular, to help the student transfer the understanding of Boolean IR from classic IR to the database context, we have provided a very similar exploration interface where elementary queries are constructed not from keywords as in classic IR but from restrictions on various fields of a database record (i.e., year = 2000 and publisher ! = "O'Reilly"). Technically, the tool is implemented as a Java servlet working on a dedicated server.

Interactive Visualization of Vector Information Retrieval Model

The vector IR model is different from

```
Brown Dog Jumps Dog Lazy Dog Fox Fox Fox
Over Over Brown Lazy Quick Dog Brown Brown
Dog Lazy Brown
```

Figure 2. A sample document used by the vector model visualization tool. On Figure 4 this document is marked as D2.

the Boolean model in many aspects. While the query in this model is also a set of terms (keywords), the terms can be weighted, stressing their relative importance. To perform the matching process, the query is converted into a weighted vector of terms. Similarly, all documents are represented as weighted vectors of terms. The weight of a specific term in a document vector represents the collection-adjusted importance of this term in the document content. The matching process is based on vector algebra. The goal of vector matching is to produce a relevance value for each document, which reflects how similar this document is to the query. The relevance value is produced by calculating the Euclidean or angular distance between the query and document vectors. Once relevance values are calculated, all documents can be ordered by their relevance to the query. Vector matching produces a ranked list, not just a subset of documents as Boolean matching does.

The hardest component of the vector matching process to understand is the calculation of the relevance value. While the geometrical nature of the relevance measure (Euclidean or angular distance) is relatively straightforward, the actual relevance values are produced by performing many operations with components of very large vectors. Not surprisingly, the final results of these calculations are much less evident to students than the results of Boolean operations with document sets. As we observed in the classroom, it is frequently hard for students to understand why a specific document was ranked particularly high or low in the resulting list.

As in the case of the Boolean visualization tool, the goal of the vector matching visualization was to uncover the steps of the matching process, which are hidden from the users of vector-based search systems. While these systems never go further than displaying the final relevance value of retrieved documents, we wanted to show how this relevance is calculated in the vector model. To uncover this process, our tool offers a small collection of documents formed from just 7 different terms (so that query and document vectors are very short). Figure 2 shows a sample document in this collection. The students explore vector matching by writing and executing simple vector queries, which are weighed sets of terms. A new query can be formed step by step by adding one term (selected from the same set of 7 terms) with its weight at a time. The example in Figure 3 shows a query formed by two terms: term *Dog* with weight 3 and term *Fox* with weight 1. For simplicity we use integer weights when forming a query; however, after the query is formed, its vector is normalized.

The results of the query execution are shown as two tables (on the right side of Figure 4): one table presents ranking based on Euclidean distance and the second presents ranking based on the cosine measure. The more relevant a document is to the query, the smaller the distance measure and the larger the cosine measure. To dig deeper, a student can click on any relevance value (shown in blue) and see in a popup window (bottom of Figure 4) how this value is calculated from the components of the query and document vectors. (The example in Figure 4 uses Euclidean distance.) The vectors for the

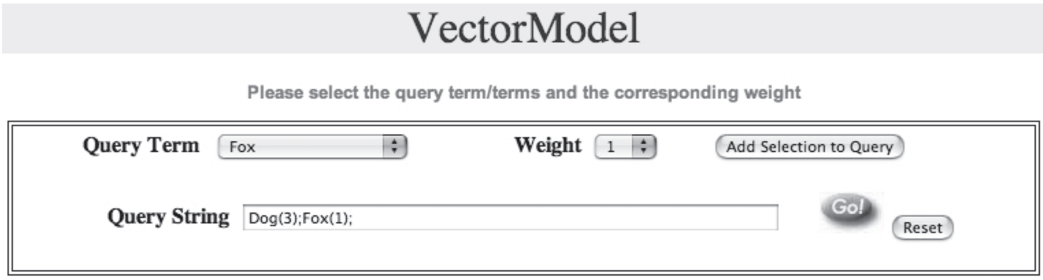


Figure 3. The process of forming a query in the vector model visualization tool.

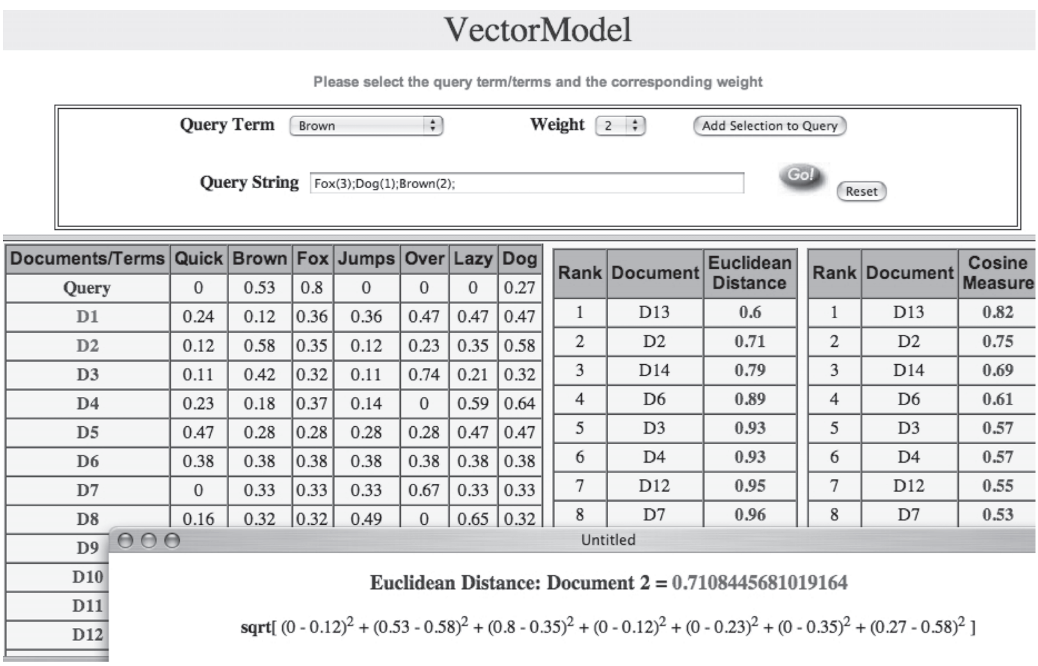
query and documents are shown on the left alongside the ranking results, so the students can see how the vector components are used in the calculations. By clicking on a document number (shown in blue), students can view the content of each document in a popup window. Thus the visualization tool makes all the traditionally hidden steps of the vector matching process visible: from documents and query to their vectors, from vectors to relevance values, and from relevance values to ranked lists.

In addition to the interactive simulation presented above, the tool also offers a brief tutorial on vector matching. It is implemented using a combination of Java servlets and client-side Javascript programs.

Classroom Evaluation of Interactive Visualization

Study Design

To examine whether our visualization tools for teaching information retrieval are effective in a real educational pro-



Documents/Terms	Quick	Brown	Fox	Jumps	Over	Lazy	Dog
Query	0	0.53	0.8	0	0	0	0.27
D1	0.24	0.12	0.36	0.36	0.47	0.47	0.47
D2	0.12	0.58	0.35	0.12	0.23	0.35	0.58
D3	0.11	0.42	0.32	0.11	0.74	0.21	0.32
D4	0.23	0.18	0.37	0.14	0	0.59	0.64
D5	0.47	0.28	0.28	0.28	0.28	0.47	0.47
D6	0.38	0.38	0.38	0.38	0.38	0.38	0.38
D7	0	0.33	0.33	0.33	0.67	0.33	0.33
D8	0.16	0.32	0.32	0.49	0	0.65	0.32
D9							
D10							
D11							
D12							

Rank	Document	Euclidean Distance
1	D13	0.6
2	D2	0.71
3	D14	0.79
4	D6	0.89
5	D3	0.93
6	D4	0.93
7	D12	0.95
8	D7	0.96

Rank	Document	Cosine Measure
1	D13	0.82
2	D2	0.75
3	D14	0.69
4	D6	0.61
5	D3	0.57
6	D4	0.57
7	D12	0.55
8	D7	0.53

Figure 4. The results of query execution produced by the vector model visualization tool. The popup window at the bottom shows how the Euclidean distance value of 0.71 was produced for document D2 given the document and query vectors shown above.

Table 1: Classroom Studies of Boolean and Vector Model Visualization Tools.

Semester	2001 Fall	2002 Fall	2004 Summer	2004 Fall	2005 Fall	Total
Number of students	26	27	10	13	19	**95**

cess, we conducted several studies with students of information retrieval courses. In each of these studies we evaluated several visualization tools. This section reports results of five classroom studies performed in the context of a graduate course, "Information Storage and Retrieval," offered annually at the School of Information Sciences, University of Pittsburgh. These studies were performed between Fall 2001 and Fall 2005. The Boolean and vector model visualization tools, which are the focus of this paper, were evaluated in each of the five studies. Table 1 shows the number of students involved. During 5 semesters, a total of 95 students completed the study, with a mean of 19 students per class (SD = 7.58).

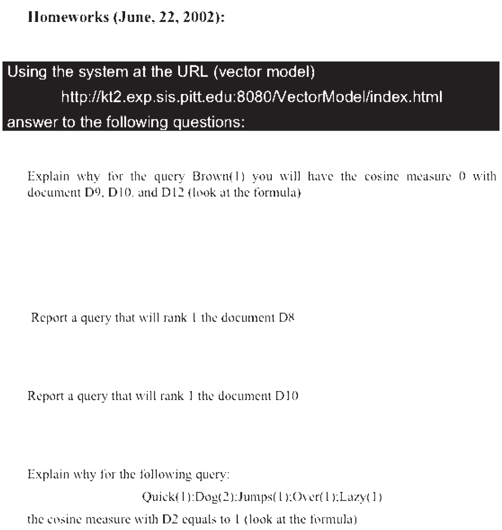
All studies of our visualization tools have the same multi-stage design presented in Figure 5. The goal of this design was to assess both objectively and subjectively whether the visualization tools are actually working as useful aids to help the students better understand the underlying concepts. As an objective measure, we used knowledge gain from pre-test to post-test. As a subjective measure, we used student answers to a questionnaire about the systems and their features. The studies of different tools were performed during different weeks of the semester. In a given week, the students learned the concepts of the subject (i.e., Boolean or vector model) during a lecture and the instructor introduced the corresponding visualization tool, encouraging them to utilize it for their deeper understanding of the knowledge they acquired in the classroom.

At the end of the lectures devoted to the Boolean and vector space models, the students took a pre-test and were given homework assignments which, among other things, specifically required them to use the tools introduced in the lecture for actively solving hands-on problems. The role of this homework was to engage the students in *active exploration* of the visualization tools in the context of problem solving. Figure 6 shows an example of the homework on the vector space model.

To evaluate the results of their learning, the students took a post-test before the next lecture (exactly one week from the first lecture on the topic), and filled in a questionnaire about the tools and their impact. The post-test was identical to the pre-test. The questionnaire included 14 questions, seven for each tool (Table 2). A five point Likert scale was used for the questions ranging from 1 (Strongly Agree) to 5 (Strongly Disagree).

The role of the questionnaire was to evaluate student attitudes to the visualization tools and their features. However, the questionnaire alone was not sufficient for a reliable evaluation. To make sure that the student's positive (or negative) attitude is grounded in the student learning experience, we had to check whether student knowledge increased after using this tool, and examine the correlation between *knowledge gain* and attitude.

Figure 5. The classroom study procedure.


Homeworks (June, 22, 2002):

Using the system at the URL (vector model)
http://kt2.exp.sis.pitt.edu:8080/VectorModel/index.html
answer to the following questions:

Explain why for the query Brown(1) you will have the cosine measure 0 with document D9, D10, and D12 (look at the formula)

Report a query that will rank 1 the document D8

Report a query that will rank 1 the document D10

Explain why for the following query:
Quick(1):Dog(2):Jumps(1):Over(1):Lazy(1)
the cosine measure with D2 equals to 1 (look at the formula)

Figure 6. Homework assignment example.

Knowledge increase (positive knowledge gain) after using the tool can support the student's positive attitude and provide additional evidence in favor of the educational effectiveness of the tool. In contrast, the lack of knowledge increase (or knowledge decrease) would cause us to question the educational value of the tools even in the presence of positive feedback. Note that the format of a classroom study does not allow us to use knowledge gain alone as a reliable indicator of the effectiveness of the tools. While we attempted to minimize student learning from other sources by placing the pre-test after the lecture presentation of the corresponding topic, we were not able to prevent students from using other sources of knowledge during their work on the homework assignment. Only a controlled lab study could assure that the registered knowledge gain was the result of student work with the tools. Given that our main goal was to explore the value of the tools in a real educational process, we accepted the shortcomings of the knowledge gain measure and applied it as a secondary indicator.

To calculate *knowledge gain*, which measures an increase in the student's knowledge of the topic, we used pre- and post-test scores [Equation (1)]. Because of the minor differences in the number of pre/post-test questions over the study semesters, we used normalized pre- and post-test scores (both were normalized to range from 0 to 10) to calculate the knowledge gain. This allowed us to evaluate the effect of the tools over all five semesters.

$$\textit{Knowledge Gain} = \textit{Normalized Post Test Score} - \textit{Normalized Pre Test Score} \qquad (1)$$

With two evaluation measures engaged, we can formulate two formal hypotheses to assess whether our Web-based interactive visualization tools are educationally effective.

Table 2: Survey Questions for the Subjective Feedback Analysis.

Number	Question
Q1	Before I used this simulation, I understood the Boolean/vector model very well.
Q2	The visual nature of the system helped me to understand the Boolean/vector IR model better.
Q3	The interactive nature of this system (in contrast with a set of static pictures) is very important for learning.
Q4	The interface for the system was easy to understand.
Q5	I think that in the context of the Information Retrieval course the system should become one of the key course tools.
Q6	I would recommend the system to a friend who is taking an IR course next semester.
Q7	The system should be used in teaching information retrieval.

Table 3: User Feedback on the Boolean Model Tool.

Semester	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Strongly Agree	0.297	0.317	0.422	0.391	0.125	0.234	0.250
Agree	0.500	0.476	0.469	0.438	0.484	0.609	0.641
Neutral	0.125	0.190	0.109	0.141	0.344	0.141	0.063
Disagree	0.063	0.016	0.000	0.031	0.047	0.016	0.047
Strongly Disagree	0.016	0.000	0.000	0.000	0.000	0.000	0.000

H1: The student will feel that the visualization tools are helpful in understanding crucial concepts of information retrieval.

More specifically,

H1-1: The subjects will answer positively to the questions asking about the effectiveness of the visualization tools.

H2: Student knowledge will increase after their work with the tools.

That is,

H2-1: The knowledge gain calculated as a difference between post- and pre-test scores will be positive.

The following subsections report the results of the data analysis performed to evaluate these two hypotheses for each of the visualization tools.

Subjective Feedback Analysis

This section analyzes students' subjective feedback on the visualization tools. Seven questions per model were asked each semester (some semesters also included one additional question). Students' responses indicated a value from 1 to 5, where 1 is the most positive answer to each question. The distribution of student answers is shown in Tables 3 and 4 and Figures 7 and 8.

Q1 differs from the rest of the questions since it is focused on the students'

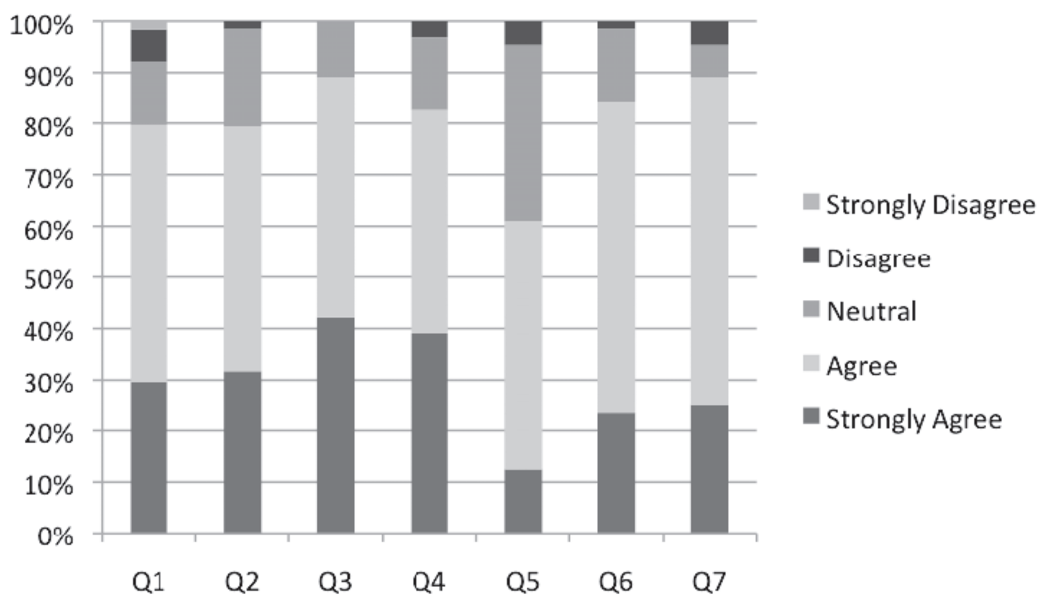


Figure 7. User feedback for the the Boolean model visualization tool.

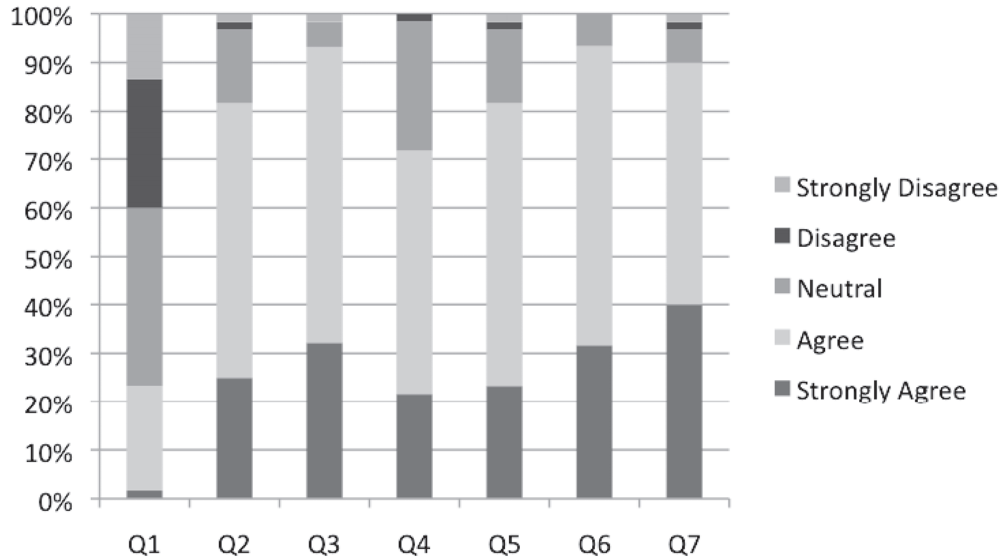


Figure 8. User feedback for the vector space visualization tool.

prior knowledge about the corresponding information retrieval model. As the data shows, about 80% of the subjects agreed or strongly agreed that they understood the Boolean model. In contrast, only 23% agreed or strongly agreed that they understood the vector model. This result is consistent with the objective analysis in the previous section, where the average pre-test score of the Boolean model was 7.87 out of 10.0 whereas that of the vector model was only 4.58.

The next two questions (Q2 and Q3) assessed the usefulness of the visual and interactive nature of the tools in the classroom. About 80% to 90% of the students gave positive answers (Strongly Agree or Agree) for both the Boolean and the vector models. Almost no negative answers were submitted. Moreover, the interactive nature of the system was valued even higher than its visual nature. This is strong evidence in favor of interactive visualization.

Question 4 asked how easy it was to understand the interface of the visualization tools. Overall, the students showed a high level of satisfaction with the Boolean tool interface (83%) and slightly lower satisfaction with the interface of the vector model tool (72%). This indicates that the latter interface may need further improvement.

Questions 5 to 7 asked in different words about the importance and usefulness of the tools in the context of an information retrieval course. Students' answers confirmed that the vector model tool, which helped them with a less well known topic, was most valuable for them.

Table 4: User Feedback on the Vector Space Model Tool.

Semester	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Strongly Agree	0.017	0.250	0.322	0.217	0.233	0.317	0.400
Agree	0.217	0.567	0.610	0.500	0.583	0.617	0.500
Neutral	0.367	0.150	0.051	0.267	0.150	0.067	0.067
Disagree	0.267	0.017	0.000	0.017	0.017	0.000	0.017
Strongly Disagree	0.133	0.017	0.017	0.000	0.017	0.000	0.017

Table 5: A Comparison of Test Scores for the Two Models.

	Boolean Model		Vector Model	
	Mean	Standard Deviation	Mean	Standard Deviation
Pre-test	7.87	2.38	4.58	2.23
Post-test	8.46	2.17	6.38	2.78

Over 80% of the students agreed that the vector tool should become one of the key course tools and over 90% wanted to recommend the system to students taking the course next semester. Corresponding numbers for the Boolean tool (over 60% and over 80%) are lower stressing that this tool was seen as less valuable. This may be caused by the simplicity of the Boolean model, which was relatively well understood by most the students before the start of the course. Yet, the feedback for both tools is overwhelmingly positive: both tools were considered important and useful by a clear majority of students. Moreover, about 90% of the students agreed or strongly agreed that *both tools* should be used in teaching information retrieval courses.

Knowledge Gain Analysis

As described earlier, we measured the pre- and post-test scores for two educational visualization tools for information retrieval: one for the Boolean and one for the vector model. Table shows the basic statistics comparing the scores. We can first observe that there are increases between pre- and post-test scores for both models. Test scores for the Boolean model increased by 0.59 and scores for vector model increased by 1.80. These differences were statistically significant (paired t-test) for both the Boolean and vector model tools ($p = 0.03$ and $p < 0.01$ respectively). It means there was actual knowledge gain after the students used the visualization tools.

It is interesting to observe that the starting knowledge of the Boolean model was relatively high. In this context it was a challenge for the tools to further increase the student knowledge. Indeed, the increase was quite moderate, although statistically significant. In contrast, the starting knowledge of vector model was relatively low—less than 50%. In this context, the visualization tools helped to achieve a more impressive knowledge gain, almost a 40% increase from the level of knowledge reported by the pre-test. This allows us to hypothesize that interactive visualization tools could be especially useful for teaching hard-to-understand topics, where regular educational means may be less effective.

It was also interesting that the standard deviation of pre-test scores were just 2.38 and 2.23 for the Boolean and vector model respectively. These values are relatively small, suggesting the starting levels of student knowledge within each topic were quite similar. Since this data was collected over five semesters of teaching the course, it could be used as a reliable estimation of student starting knowledge when preparing educational visualizations and other pedagogical material for the course.

Subjective Feedback versus Knowledge Gain

In the previous sections, we saw that there was an increase in students' knowledge after they took the information retrieval classes, and that they supported the use of visualization tools in the classes. That is, there was an objective effect on their knowledge and the source of

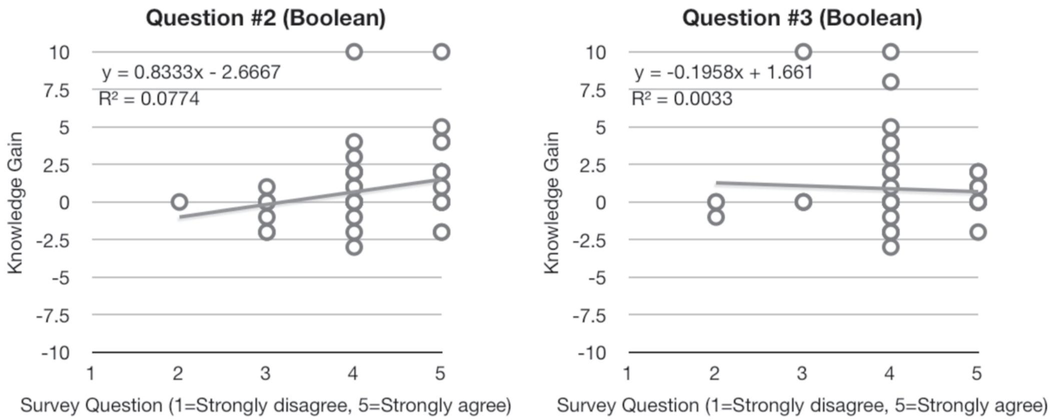


Figure 9. Comparison of subjective user feedback against knowledge gain for Boolean tools.

the effect was assumed to be the visualization tools, based on their positive response to the questions asking them about the usefulness of the tools. To confirm this assumption, we examined the relationship between the objective knowledge gain and the subjective answer from the students. We chose questions #2 and #3 from the questionnaires, which directly asked about the usefulness of the visual (Q #2) and interactive nature (Q #3) of the tools. The answers to the questions were compared with student knowledge gain scores for the Boolean and the vector model separately. Figures 9 and 10 show the results of these comparisons.

Here, three out of four graphs show a positive relationship between the subjective answers of the students and their actual knowledge gain. The students with higher knowledge gain scores tend to agree more with the statement on the usefulness of the visual nature of the Boolean tool (Figure 9 left). For the statement on the interactive nature of the Boolean tools, the effect is less pronounced and reversed (Figure 9 right). In terms of the vector visualization tools, the students with higher knowledge gain scores tend to be more positive about the visual and the interactive nature of the tools, although the effect is weaker than in the previous case (see Figure 10).

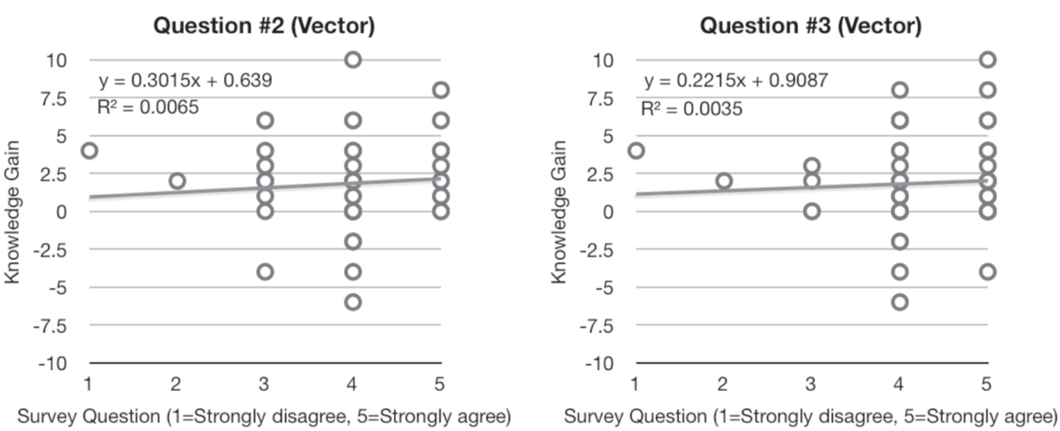


Figure 10. Comparison of subjective user feedback against knowledge gain for vector tools.

Overall, it can be observed that students who learned more (higher knowledge gain) reacted more positively to the *visualization* tools for both models—Boolean and vector.

Discussion

This paper presents our experience in developing and using a set of interactive visualization tools for teaching and learning information retrieval. By using the power of interactive visualization, our tools allow a teacher to introduce important models of information retrieval in a way that cannot be replicated by traditional whiteboard and slides, and with almost no preparation time (simply by deciding which examples to show to cover the main set of ideas). The instructor can easily accommodate very different audiences by adjusting the number of examples to show, the speed, and the granularity of presentation.

Using these interactive visualizations as learning tools further increases their value. They let the students switch from passive learning-by-reading to active and interactive exploratory learning. By exploring a number of different examples with interactive visualization tools they get a chance to achieve a better understanding of complex IR topics. The results of our classroom studies indicate that use of interactive visualization in a homework context can result in significant growth of knowledge. The vast majority of the students recognize the value of interactive visualization and recommend its use in the context of information retrieval courses. The study also demonstrated that visualization focusing on less known and harder to understand topics causes a larger growth in knowledge and is perceived as more useful. This result suggests placing higher priority on the development of visualization tools for harder to understand topics

Our results stressed the importance of developing, collecting and sharing tools for teaching information retrieval. This is consistent with the culture of sharing in IR research. A number of groups maintain Web pages of IR resources such as search software, text processing utilities, evaluation packages and topical bibliographies. We argue that similar collections of IR educational resources should be established and maintained. The Web makes it possible for researchers and educators to make their contributions to research and teaching available to others, making it much easier for others to implement. We have contributed to this process by developing a range of interactive visualization tools for teaching and learning information retrieval. We have demonstrated that these visualization tools have value for students learning the concepts behind specific information retrieval models. We have made these freely available to any IR educators or students wishing to use them on our IR resources page (http://ir.exp.sis.pitt.edu/res2/resources.php) along with similar educational resources developed by other teams. We welcome readers who wish to try these tools and hope that more educational repositories like ours will be established in the near future.

In our own future work we plan to continue exploration of interactive visualization for information retrieval. We hope to increase the number of visualization systems available for use in information retrieval courses and to perform additional formal studies of these tools.

Acknowledgements

Most of the tools in the original suite of learning environments were developed under the guidance of the first author by MSIS students of the School of Information Sciences at the University of Pittsburgh. Our School is fortunate to have many bright students and the author thanks them for their devotion and creativity. Special thanks go to Olena Scherbinina, Archana Sulebele, and

Anand Banvasi, the developers of Boolean and vector model visualization tools presented in this paper.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* New York, NY: ACM Press.

Brusilovsky, P. (2002, June 24–29, 2002). *Web-based interactive visualization in an information retrieval course.* Paper presented at the ED-MEDIA'2002—World Conference on Educational Multimedia, Hypermedia and Telecommunications, Denver, CO.

Butler, J. E., & Brockman, J. B. (2001). A Web-based learning tool that simulates a simple computer architecture. *SIGCSE Bulletin—inroads, 33*(2), 47–50.

Domingue, J., & Mulholland, P. (1998). An effective Web based software visualization learning environment. *Journal of Visual Languages and Computing, 9*(5), 485–508.

Evans, K., Yaron, D., & Leinhardt, G. (2008). Learning stoichiometry: A comparison of text and multimedia formats. *Chemistry Education: Research and Practice, 9*, 208–218.

Gordin, D. N., & Pea, R. D. (1995). Prospects for scientific visualization as an educational technology. *Journal of the Learning Sciences, 4*(3), 249–279.

Haajanen, J., Pesonius, M., Sutinen, E., Tarhio, J., Teräsvirta, T., & Vanninen, P. (1997). *Animation of user algorithms on the Web.* Paper presented at the VL '97, IEEE Symposium on Visual Languages.

Hundhausen, C. D., Douglas, S. A., & Stasko, J. T. (2002). A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing, 13*(3), 259–290.

Korfhage, R. R. (1997). *Information storage and retrieval.* New York, NY: Wiley.

Levy, R. B.-B., Ben-Ari, M., & Uronen, P. A. (2003). The Jeliot 2000 program animation system. *Computers and Education, 40*(1), 1–15.

McClean, P., Johnson, C., Rogers, R., Daniels, L., Reber, J., Slator, B. M., et al. (2005). Molecular and cellular biology animations: Development and impact on student learning. *Cell Biology Education, 4*(2), 169–179.

Naps, T., Rößling, G., Anderson, J., Cooper, S., Dann, W., Fleischer, R., et al. (2003). Evaluating the educational impact of visualization. ACM SIGCSE bulletin, 35(4), 124–136.

Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C., et al. (2006). PhET: Interactive simulations for teaching and learning physics. *The Physics Teacher, 44*(1), 18–23.

Rößling, G., Schüer, M., & Freisleben, B. (2000). *The ANIMAL algorithm animation tool.* Paper presented at the 5th annual SIGCSE/SIGCUE ITiCSEconference on Innovation and technology in computer science education, Helsinki, Finland.

Tung, S.-H. S. (1998). Visualizing Evaluation in Scheme. *Lisp and Symbolic Computation, 10*(3), 201–222.