SIGIR 2003 Workshop Report: Implicit Measures of User Interests and Preferences

Susan Dumais
Microsoft Research
sdumais@microsoft.com

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

Krishna Bharat
Google, Inc.
krishna@google.com

Andreas Weigend
Amazon, Inc.
aweigend@amazon.com

Introduction

The goal of the workshop on *Implicit Measures* was to explore how various implicit measures of user interest can be used in information retrieval and filtering applications. In most information retrieval or filtering applications, it is difficult to get explicit feedback from users about the relevance of the results, the appropriateness of the presentation, and more generally about the quality of their experience. Yet explicit judgments are assumed by researchers for many activities like the tuning and selection of ranking algorithms, information combination, user modeling, information presentation, etc. This workshop explored how implicit measures of user interest (such as dwell time, click through, and user activities like annotation, printing, and purchasing) can be used to develop predictive models for a variety of purposes. In the context of information retrieval such models could be used to: improve ranking and relevance assessment (e.g., the extent to which implicit measures can be used to evaluate the quality of systems, ranking algorithms and recommendations, or as input to relevance feedback algorithms); personalize search, filtering or presentation; personalize considering both individual and aggregate data; adaptive link generation or web site design; etc. An examination of theoretical issues such as modeling approaches (Bayesian techniques and other predictive models), gold standards for user behavior (e.g., relevance judgments, purchases), combining implicit and explicit preferences, and biases introduced by reliance on implicit measures were also encouraged.

Since this was the first workshop on the topic at SIGIR, we encouraged participation from people with different backgrounds and perspectives including theoretical modeling, experimental analysis, and applications development. Applications from information retrieval, collaborative filtering, e-commerce, user modeling, and human-computer interaction were encouraged. Participation from both academia and industry was encouraged. As an outcome of the workshop we hoped to identify key theoretical modeling issues, systematize engineering principles and best practices, and spark new research directions.

The workshop was held on August 1, 2003 in Toronto, Canada, immediately following the SIGIR 2003 conference. The workshop drew about 35 participants, representing a nice mix of academia and industry, and information retrieval and learning. To encourage interaction among participants, we limited the number and length of the talks to 15

minutes. This was followed by a 5 minute discussion by one of the organizers and additional open discussion among participants. This seemed to work well as a way to encourage interaction. We also had poster presentations available during the lunch break, which worked well, although only two posters were presented.

Presentations

Below we provide a brief summary of the workshop presentations to give a sense for the range of themes and topics covered. The complete workshop abstracts and presentations are available at:
http://research.microsoft.com/~sdumais/SIGIR2003/SIGIR2003-ImplicitWorkshop.htm

The workshop started with an invited talk by Steve Lawrence (Google Inc.), *Implicit feedback: Good may be better than best.* This talk reported on experiences from the CiteSeer and NewsSeer projects that Steve led while he was at NEC. Steve started with the high-level observation that "good may be better than best" for many applications. The web has many limitations (e.g., dead links, lack of support for author royalties, inconsistent content and linking, no version control, etc.), but it is extremely simple to use. Better system designs exist (e.g.,Xanadu), but are largely unused in part because they impose added complexity for system design and maintenance and extra effort on creators and users of information. CiteSeer is an automatically generated virtual library and search system for scientific literature, consisting of more than 600,000 documents. Because it is automatically generated, it is easy to create and scale, but it has some errors. Both implicit and explicit feedback is collected. More than 300,000 explicit corrections were received from authors who want to correct errors or update their citations. Only 0.17% of document accesses were followed by explicit feedback, yet every access and citation provide implicit measures of interest. High citation papers are downloaded and accessed more often than low cited papers, but they receive the same rate (both low) of explicit ratings. User profiling and paper recommendation were also available for a while, based on a combination of implicit and explicit information. Different recommendation algorithms were evaluated using number of recommendations viewed, as the measure of success. Similarly, NewsSeer provided news recommendations based on implicit measures, which could be supplemented with explicit ratings. In several applications, implicit feedback was more useful than explicit feedback because there was much greater participation. Implicit feedback can be encouraged by clever interface and system design. Explicit feedback can be obtained when the incentives are high enough and input is easy enough.

Bill Grosky (University of Michigan-Dearborn) talked about *Deriving emergent web page semantics*. This work described how browsing paths can be used to find the latent semantic structure of multimedia web pages. The semantics of a page is richer than can be defined by the page's authors. Explicit links are augmented by studying user browsing behaviors. Observing browsing patterns allows one to discover different meaning for the same document (by looking at differences across users), and similar documents (those visited by the same person in a short period of time). Each path is represented by a vector which captures textual, visual and structural keywords from the pages in the path. Latent semantic analysis is performed on the matrix to identify "semantic breakpoints" which are used to identify semantically coherent subpaths.

David Brown (Worcester Polytechnic Institute) talked about, *Curious Browsers: Automated gathering of implicit interest indicators by an instrumented browser.* Brown, Claypool and their students have developed an instrumented browser (the Curious

Browser) in order to study the relationship between various implicit measures of interest and explicit user judgments of page quality. Two experiments, with college students browsing in a computer laboratory were reported; 80 students participated in each experiment, contributing judgments on 2500 and 1000 pages in the two experiments. Box plots (and Kruskal-Wallis tests) were used to identify implicit measures that were strongly related to explicit page ratings. Time on page, combined scrolling time, and mouse movement (horizontal and total movements) were correlated with explicit interest. Mouse clicks and various keyboard activities were not correlated with explicit judgments. The data sets from these experiments are available from the authors.

Steve Fox (Microsoft) talked about work on, *Evaluating implicit measures to improve the search experience*. The goal of this work is similar in spirit to the previous Curious Browser work, but there were several important new contributions. First, the task of interest was the extent to which implicit measures could predict explicit ratings of web search results, at both an individual page and search session level. Second, an instrumented browser was used by almost 150 people in their normal work environment for several weeks. Finally, Bayesian predictive modeling and novel sequence analysis techniques were used to analyze the results. The WinMine toolkit was used to build Bayesian models to predict page and session judgments of Satisfaction, Partial Satisfaction and Dissatisfaction. Dwell time, position, scroll count, and exit type (how the user exited a page) were predictive of individual page judgments; dwell time, number of results sets and exit type (how the user exited a session) were predictive of session judgments. Printing and adding to favorites occur infrequently but are highly indicative of satisfaction when they occur. Combined measures better predicted satisfaction than clickthrough alone. Sequence analyses, that they call 'gene sequences', are a promising new technique for describing a series of user interactions within a session.

Doug Oard (University of Maryland), Anton Leuski (University of Southern California), and Stuart Stubblebine (Stubblebine Research Labs) talked about, *Protecting privacy of observable behavior in distributed recommender systems*. When collecting any implicit (or explicit) measures, protecting the privacy of users is an important consideration. This group focused on the problem of privacy in the context of distributed systems. They assume that guarantees about privacy protection could yield behavior-based evidence about the utility of objects in a variety of applications. They outlined a framework for thinking about observable behaviors and initial thoughts on how to architect privacy protection into recommender systems. Four behavior categories (examine, retain, reference and annotate) and three scopes (segment, object, class) were identified and common implicit measures categorized. E.g., bookmarking and purchasing are retain activities on objects. There are some observable public behaviors (such as hypertext linking, publications and citations), but architectural assurances of privacy could increase the availability of implicit measures. An Item, Behavior, Feature, Recommendation (IBFR) architecture and cryptographic ideas for protection were outlined. This talk generated lots of discussion about centralized ratings services, especially from a business perspective (i.e., will Amazon be willing to share data with Barnes and Nobel). It was great food for thought, which led us into our lunch break.

During the lunch break, two posters were available for viewing and discussion. Diane Kelly (Rutgers University) showed her work, *Understanding implicit feedback: A naturalistic user study*. The poster described a study in which a wide range of implicit measures were collected in a realistic usage setting -- 7 users were given laptop

computers and printers, and their activities monitored for a 14 week period of time using a combination of computer instrumentation, surveys and detailed interviews.

Thorsten Joachims (Cornell University) described his work on, *Evaluating retrieval performance using clickthrough data*. The goal of this work is to learn retrieval functions using implicit behavior. There are numerous pitfalls of measures such as number of clicks or rank of clicks. Relative preference orderings (e.g., if a user skips link a and clicks on link lib then the user preference reflects rank(b) < rank(a)) provide useful training data for learning. A support vector machine method was proposed for learning ranking functions and evaluated for learning meta-search parameters.

Jie Wu (Swiss Federal Institute of Technology, EPF) spoke about work on, *Semantic web graph implied by user preferred activities*. This work reported on preliminary experiments looking at semantic web graphs (based on user interaction patterns), rather than static web graphs based on connectivity. The analytical approach was inspired by recent work in swarm intelligence in which global optimization is performed using distributed local computations. The rankings generated by graph characteristics based on usage indeed differ from those generated by a static connectivity analysis, but the extent to which they are better is an area for future research. The extent to which the swarm intelligence ideas should be taken as more than an analogy led to some interesting discussion.

Travis Bauer (Sandia National Laboratories), *WordSieve: Learning task differentiating keywords automatically*. This work looks at techniques for automatically identifying a user's task context. An unsupervised competitive learning technique was developed to extract context-differentiating words from document streams automatically. Terms compete with one another in a local neural network learning technique. This real-time algorithm requires only local information, and can thus adapt quickly to evolving user interests. The highest scoring terms can then be used to characterize the topic of current user activity. The WordSieve system was evaluated using a small set of web browsing data and shown to be comparable to LSI matching and better than tf*idf and log-entropy vector matching, all of which require global corpus statistics.

Antonio Bahamonde (Universidad de Oviedo at Gijón), *A clustering algorithm to find groups with homogeneous preferences.* The goal of this work is to learn people's preferences. The work was done in the context of food preferences, but can be generalized to information access problems where ranking is important. In order to learn ranking functions, it is desirable to cluster preference ratings of different users. A preference clustering algorithm for pair wise preferences was developed and evaluated using the EachMovie database.

Andreas Lorenz and Andreas Zimmermann (Fraunhofer Institute for Applied Information Technology) were scheduled to present a talk on, *Inferring interests from user movements: The LISTEN approach,* but had travel problems and were not able to attend. This work is based on an interesting application in which they developed adaptive audio presentation for a museum exhibit. The extended abstract can be viewed at the workshop web page.

The workshop ended with a final open discussion and wrap-up, which is summarized below.

Reflections

Perhaps the most salient characteristic of the workshop was the tremendous diversity in the application areas (information retrieval, web browsing, topic identification, preference analysis), implicit measures considered (links, citations, dwell time, scrolling, viewing) and analysis techniques developed (correlation, Bayesian modeling, completive neural networks, support vector machines, singular value decomposition, clustering). The issue of privacy cut across all talks, but this was one of the few common themes. Given the tremendous diversity of work in the general area of implicit measures, it seems premature to identify key issues and best practices.

One suggestion was further broaden the perspectives considered by reaching out to other communities. Similar workshops could be held at e-commerce, machine learning, data mining, human-computer interaction or artificial intelligence conferences. Participants will look into this and follow-up as appropriate. Although many measures were considered, some were notable by their absence, such as eye movements, purchasing and other commerce activities, and any systematic coverage of individual or group differences. Personalization was an area that people thought was important to pursue.

Another suggestion was to narrow the focus to a topic, such as web search or e-business, in order to better understand the merits of different approaches on a common problem and to outline key research challenges. One of the difficulties in doing this is the lack of a common data set. There is a small collection available from the Curious Browser experiments. A larger data set is desirable. Another challenge was simply gathering the data for additional applications of interest. Client-side instrumentation was especially challenging, and some groups will pursue discussions of sharing browser and system instrumentation code.

Two attendees, Diane Kelly (Rutgers) and Jaime Teevan (MIT), put together an annotated bibliography of work on implicit measures used in information retrieval applications. They characterized the research in terms of Oard and Kim's implicit behavior taxonomy. This bibliography appears in the SIGIR 2003 Forum, Fall Issue.

Acknowledgements

We would like to thank the SIGIR conference organizers (Charlie Clarke and Gordon Cormack) and the local arrangements chair (Elaine Toms) for their excellent support of the workshop.