Search engines that learn from experience

Some components of search engines that rank search results need periodic "tune-ups" when the environment changes. An exception is the search engine Osmot, developed by CS professor Thorsten Joachims and his student Filip Radlinski. When fielded on Cornell's Library Web pages, Osmot tuned itself to this new collection and user base without expert intervention. To avoid making the same mistake twice, Osmot observes how users react to results and uses machine-learning to update its ranking function. For example, it quickly learned that users with the query "oed" wanted to visit the library gateway to the Oxford English Dictionary, even though this page does not contain the word "oed".

CS professor Thorsten Joachims: When users reformulate queries and select links, their decisions provide the search engine with training data. The system can leverage user skills to improve itself.

Osmot is an example of how collaborations between CS and Cornell's new Program in Information Science combine math topics traditionally pursued in computer science with research in human factors. "You can learn a lot about people by watching how they act and react," says Joachims. "When users reformulate queries and select links, their decisions provide the search engine with training data. The system can leverage user skills to improve itself."

Eye-tracking experiments have provided interesting insights in other computer science areas. In the late 1990's, Eric Aaron, CS Professor David Gries's PhD student, performed eye-tracking experiments with Professor Spivey of Psychology to analyze how students developed calculational proofs. The findings confirmed some expected behaviors, based on strategies and principles taught in the Gries-Schneider text *A Logical Approach to Discrete Math*, and uncovered other interesting patterns, such as the tendency to attend to particular premises despite their not being used in the proof under consideration.

The most intriguing property of using observable user behavior as implicit feedback is its availability. It directly reflects individual preferences, and it can be gathered without user effort. So, in principle, search engines need not be one-size-fits-all; instead, they can learn what each user means with their queries. For example, the word "keyboard" in a query from a user at cs.cornell.edu is less likely to refer to a musical instrument than for an average user. A search engine that knows its users from their reactions to the results of previous searches can make better guesses about the meaning of future queries and documents. The better the guess, the better the retrieval quality.

It is not always clear how to interpret user behavior reliably. For example, does a click on a link in the search results really mean that the link is relevant? The answer is "no", says Joachims, who investigated the question with CIS professor Geri Gay and research associate Helene Hembrooke of Information Science, along with postdoc Bing Pan and grad student Laura Granka. They used eye-tracking experiments to analyze the decision process of search-engine users. Other factors influence clicking behavior, it turns out, most prominently the position in the ranking. "In Google We Trust," said Hembrooke. "Whenever we moved a link to the top of the ranking, it received more clicks." While clicks do not indicate relevance on an absolute scale, other interpretations of clicks do give highly accurate feedback. For example, if a user does not click on the top link but instead reformulates the query and clicks on a link there, then, with high probability, the clicked link is more relevant than the top link of the original query.

"We learned in these studies that we can get accurate relative preferences between links but no absolute relevance judgments," says Joachims. However, most traditional machine learning algorithms can use only absolute feedback. To overcome this problem, they adapted the Support Vector Machine learning method to make use of relative feedback. Here, research on human factors in search-engine use uncovered and directed the need for research on machine learning methods.

The next challenge is to scale these methods to collections of the size of the Web. Retrieval functions that explicitly model all users and sites on the Web will be among the largest machine learning problems ever attempted, involving billions of features and millions of examples every day. But, given that tractability was pushed from hundreds of features and examples 15 years ago to hundreds of thousands of features and examples today, such problems are no longer beyond reach.

Eye-tracking experiments are used to analyze the decision process of search-engine users.

The father of information retrieval

CS professor Gerry Salton is the man most responsible for the creation and coming of age of information retrieval (IR).

Salton published more than 150 research articles and five texts on information retrieval. His honors are too numerous to mention. Among the most prestigious are a Guggenheim Fellowship (1962), ASIS Award for Best Information Science Paper (1970), Best Information Science Book (1975), the first ACM/SIGIR Award for Outstanding Contributions to Information Retrieval (1983), the Alexander von Humboldt Senior Science Award (1988), and the ASIS Award of Merit (1989). The ACM/SIGIR Award was subsequently renamed the Gerard Salton Award. He became an ACM Fellow in 1995.

Salton *was* information retrieval. At the heart of every IR system, Web-based or otherwise, is the set of keywords and phrases that are collectively used to describe, or index, each document. In stark contrast to the standard indexing approach requiring manual assignment of index terms to texts, he was a very early and vocal proponent of *automatic indexing*. He proposed a scheme in which *every* word in a document (except for the most common ones) would be used as an index term. This type of full-text indexing technique comprises the core technology in virtually all of today's Internet search engines. Salton's subsequent work addressed, in turn, the critical components of automatic full-text indexing retrieval systems: term weighting, relevance feedback, document clustering, extended boolean retrieval, term discrimination value, dictionary construction, term dependency, phrase indexing, semantic indexing via thesauri, text understanding and structuring, passage retrieval, and even document summarization.

Salton is best known for his vector space model of information retrieval, upon which modern retrieval systems are based; and for the SMART system, his publicly available automatic text processing system, which incorporates the vector space model. SMART, which was known as *Salton's Magical Retriever of Text* (later given the dull interpretation *System for the Manipulation and Retrieval of Text* by more pedantic professors), rapidly matured to the stage where it was the most advanced information retrieval system in the world for many years. It remains a powerful experimental vehicle. Individual, a news clipping service, licensed the technology directly. WAIS (Wide Area Information Server) and DOWQUEST (a tool for the Dow Jones news wire) and others use technology derived from SMART, and many new systems have leveraged his years of research. Today, with the World Wide Web and massive digital libraries, IR has come of age.

The epitaph for Sir Christopher Wren, the architect of St. Paul's Cathedral in London, reads, "If you wish to see his monument, look around you." If you wish to find Salton's monument, use any one of the many text-based search engines for navigating the World Wide Web.

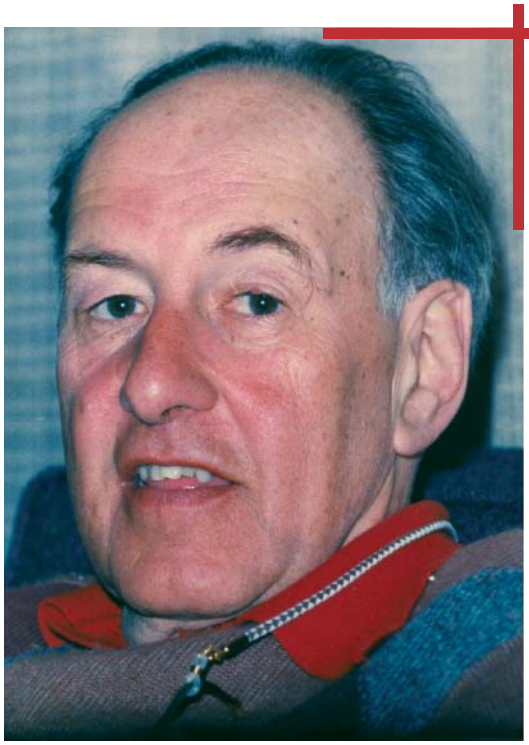
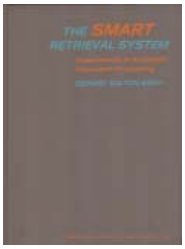
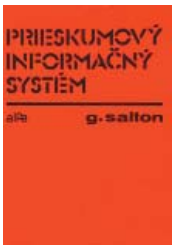


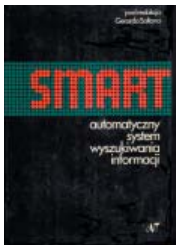
Photo: Professor Edgar Rosenberg

Gerry Salton, 1927-1995

The Salton library







Dan Huttenlocher is the CASE New York State Professor of the Year. The award covers all disciplines. It is given by the Council for Advancement and Support of Education for impact and involvement with undergraduates, scholarly approach to learning, and contributions to undergraduate education.

David Gries receives a Cornell Presidential Weiss Fellowship for his contributions to undergrad education. Three such awards are given each year; Cornell has 1600 faculty members.

T.V. Raman receives the ACM Doctoral Dissertation Award for his PhD thesis, *Audio System For Technical Readings* (Springer-Verlag, 1998). Raman's advisor was David Gries. Raman is now a researcher at Google.

Researchers Jim David, Dean Krafft, and Carl Lagoze release Dienst, which becomes the foundation for future digital library interoperability.

Eva Tardos, Joe Halpern, Jon Kleinberg join.

1995

CS mourns the passing of Gerry Salton, a founding member of the department and the father of information retrieval.

David Gries receives the ACM Karlstrom Outstanding Educator Award. The citation reads, "His visionary emphasis on critical thinking and mathematical precision has dramatically changed the face of computer science education … ."

David Gries receives an honorary doctorate from Daniel Webster College.

Fred Schneider becomes Professor-at-Large at the University of Tromso, Norway.

Juris Hartmanis receives the Bolzano Gold Medal of the Academy of Sciences of the Czech Republic for Merit in the Field of Mathematical Sciences.

Juris Hartmanis receives an honorary doctorate from the University of Dortmund.

Ken Birman chairs a DARPA ISAT study on survivability of critical infrastructure; Fred Schneider is on the committee. The study establishes a major DARPA effort in the area and lays the groundwork for a broader government engagement of the challenge.

Neil Immerman (former student of Juris Hartmanis) and Róbert Szelepcsényi get the Gödel prize for their paper showing that nondeterministic logarithmic space is closed under complement.