Web Intelligence – Mining the Web for Relevant Information: Concepts & Applications

Jaideep Srivastava

The exponential growth of the Web has made it into a huge inter-connected source of information. It is estimated that the Web today has close to 7 billion static pages, and millions of databases at various web sites that can generates many more billions of dynamic pages. Today there are three main paradigms of accessing information from the Web: (i) using a search engine like Google to find pages containing certain keywords, (ii) using a Web index like Yahoo to find pages relevant to certain concepts, and (iii) randomly browsing or "surfing " the web – going from page to page – using browsers like Internet Explorer. While these are all very powerful ways of using the web, they are designed to access information from the Web in an 'exploratory manner'. While this is useful for individuals looking for information – especially for their private purposes, the needs of organizations or corporations accessing the Web as an information source is turning out to be quite different. For example, financial analysts in the equity research department of Morgan Stanley do not care (as much) about being able to search through the 1.4 billion web pages that Google indexes, but are very interested in full details that can be obtained from the top 100 financial sites, e.g. MorningStar, Wall Street Journal, etc. While there is always some degree of exploration to discover new and useful sites, there is a far greater need to establish 'information channels' that will provide the equity researchers high-quality and detailed information from the top 100 financial sites on a continuous basis. Furthermore, the information provided is much more valuable if it in the form of 'business intelligence' rather than just text found on web pages. This automatic extraction of business intelligence from relevant web sites is being called 'Web Intelligence'.

The past couple of years have seen the introduction of a set of technologies that are making Web Intelligence possible. **First** of these is Web Content Acquisition, which provides the capability of not only searching static pages at a site, but also of crawling entire sites, and executing search engines of web sites to extract dynamic pages. There has been considerable work on developing tools that can be used for rapid development of wrappers and crawlers used for these tasks. The **second** technology is Web Mining, which is the application of data mining techniques to content, structure, and usage data from the Web to extract higher order knowledge. Web Mining has seen rapid growth in the past few years. The **third** and final technology is the creation & management of user profiles, which is being used for personalization and relevance analysis.

In this talk we will first introduce the concept of Web Intelligence, and provide evidence of its growing importance in the corporate world. Second, we will provide a survey of the key concepts in the component technologies. Next, we will discuss the architecture of a Web Intelligence system, and give examples from the commercial world. Finally we will discuss two case studies of use of such systems – one in the financial space, and another in the product design space.

Biographical Sketch

Jaideep Srivastava
srivasta@cs.umn.edu

Jaideep Srivastava received his B.Tech. from the Indian Institute of Technology in 1983, and M.S. and Ph.D. from the University of California - Berkeley in 1985 and 1988, respectively. Since 1988 he has been on the faculty of the University of Minnesota, where he is a Professor. For over 15 years he has been active as a researcher, educator, consultant, and invited speaker, in the areas of data mining, databases, artificial intelligence, and multimedia. He has established and led a database and multimedia research laboratory, which has graduated 16 Ph.D. and 36 M.S. students, and in the process published over 135 papers in journals and conferences. Throughout his career Jaideep has had an active collaboration with the industry, both for collaborative research and technology transfer.

Since 1999 Jaideep has been on leave from the University of Minnesota, during which period he has spent time at Amazon.com (www.amazon.com) as the Chief Data Mining Architect, at Yodlee Inc. (www.yodlee.com) as Director – Data Analytics, and at Kandeo, Inc. (www.kandeo.com) as the Chief Technology Officer. This wide-ranging industry experience has provided Jaideep a unique perspective on the application of various computer science technologies in the Internet economy.

Jaideep is an often-invited participant in technical as well as technology strategy forums. He has given more than a hundred talks in various industry, academic, and government forums. He has served on the program committee of a number of conferences, and is on the editorial board of various journals. The federal government has solicited his opinion on computer science research as an expert witness. He has also served in an advisory role to the governments of India and Chile on various software technologies. Jaideep is a member of the ACM, and a senior member of the IEEE.

Samples of Jaideep's public domain work are available at http://www.cs.umn.edu/Research/mmdbms/, http://www.cs.umn.edu/Research/websift/, and http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Srivastava:Jaideep.html.