
World Wide Web: Internet and Web Information Systems, 8, 99–100, 2005

Editorial

The World Wide Web has quickly emerged as a critically important system for information dissemination, retrieval, and electronic commerce. Research and development of the Web has been occurring at a pace that has rarely if ever been matched in other technological fields. The rapid development of the Web has been enabled by continuous breakthroughs in several areas including information retrieval and searching techniques, Web browsers and crawlers, languages for representing information, and security. Efficient techniques for serving Web data, caching, load balancing, and replication have allowed Web sites to handle ever increasing amounts of traffic while providing fast service and high availability.

This special issue includes four papers on research that is laying the foundations for the future of the Web. These four papers are enhanced versions of top papers from the Twelfth International World Wide Web Conference (WWW2003) held 20–24 May, 2003, in Budapest, Hungary. The four papers in this special issue were selected from 77 papers accepted by the main refereed paper track of the conference out of a total pool of 602 submissions.

The first paper, entitled "Query-Free News Search," studies the problem of automatically finding news articles on the Web relevant to television broadcast news. The approach taken by the authors is to extract queries from an ongoing stream of closed captions, issue the queries in real time to a news search engine on the Web, and post-process the top results to determine the news articles to show to the user. The authors compare seven algorithms and three post-processing techniques using a precision metric which is the percentage of relevant articles out of all articles returned. The best algorithm finds a relevant page every 16–20 seconds on average, achieves a precision of 84–91%, and finds a relevant article for about 70% of the topics.

The second paper, entitled "Active Service for Mobile Middleware," describes a dynamic service reconfiguration model to enhance service provision for wireless Web access. The model employs a proxy that is composed of a chain of service objects called *mobilets,* which can be actively deployed onto a network. This model offers flexibility because the chain of mobilets can be dynamically reconfigured to adapt to dynamic changes in the wireless environment, without interrupting service provision for other mobile nodes. The authors argue that potential computation overheads incurred by active services operating at the application level are insignificant compared to the benefits gained in providing highly adaptive applications operating in the dynamic environment of wireless links.

The third paper, entitled "On the Bursty Evolution of Blogspace," proposes two new tools to address the evolution of hyperlinked corpora: *time graphs* that extend the traditional notion of an evolving directed graph, and definitions and algorithms for *time-dense community tracking*, which crystallize the notion of community evolution. The authors

developed the tools and performed the study in the context of Blogspace, the space of *weblogs* (or *blogs*). The authors demonstrated that the Blogspace has been rapidly expanding in metrics of scale, community structure, and connectedness. The authors also discovered dense periods of intra-community link creation and found that the blogs that give rise to these communities are significantly more enduring than an average blog.

The fourth paper, entitled "Automating Content Extraction of HTML Documents," deals with the problem of how to extract important content from HTML documents while removing distracting features such as banner advertisements, unnecessary images, and extraneous hypertext links. The authors have developed a publicly available Web proxy to extract content from HTML pages known as Crunch. Crunch passes Web pages through an open source HTML parser which creates a Document Model (DOM) tree. Crunch extracts content heuristically, with heuristics customizable by an administrator or knowledgeable user. Multiple filters are used to remove extraneous content including an advertisement remover, a link list remover for removing content where the ratio of the number of links to the number of non-linked words is greater than a specific threshold, and an empty table remover for eliminating tables which do not have substantive information.

We would like to thank the many authors, program committee members, and other organizers who contributed to the success of the Twelfth International World Wide Web Conference (WWW2003).

Robin Chen, AT&T Labs–Research
Arun Iyengar, IBM Research

Guest Editors