

IBM Research has made major contributions to a wide range of Web-related areas. For example, its researchers pioneered new techniques for efficiently providing dynamic content, which enabled 100 percent availability under very high publishing requirements at some of the most highly accessed sports Web sites in the world. Research has played a key role in important Internet standards such as XML Schema, Resource Description Framework (RDF), and XHTML+Voice. In addition, researchers have made seminal contributions to important Web information retrieval problems, including the Web graph model, rank aggregation, and automated community extraction. With the continuously increasing span and complexity of the World Wide Web, IBM Research's goal is to efficiently address scale and variety in Web data applications and to further improve the performance and scalability of Web infrastructures.

MANAGEMENT OF LARGE-SCALE DATA COLLECTIONS

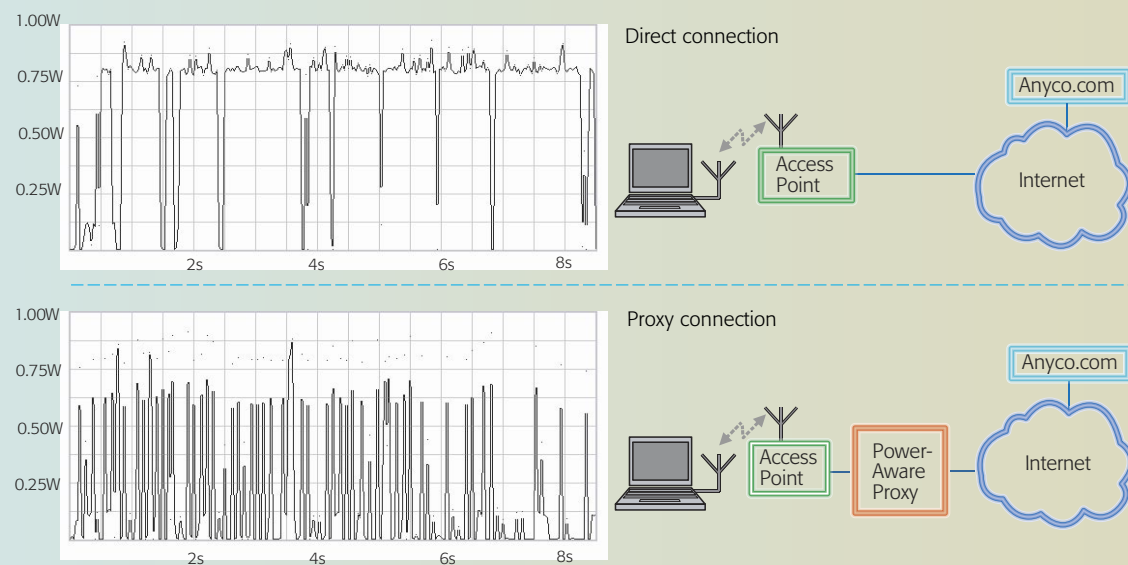
IBM research in this area addresses the modeling of large-scale, heterogeneous data collections and improving the efficiency of information retrieval over the Web and enterprise-scale data collections. A large-scale data collection, for example "everything on the Web," can be thought of as a single virtual XML unit, without actually converting and storing it in XML form. New approaches for specifying such data virtualizations are being explored along with the implementation of the XQuery language that enables users to efficiently run expressive queries.

Furthermore, information-retrieval problems related to large-scale data collections, such as those in enterprise IT infrastructures created through mergers and acquisitions, are being addressed. In these types of infrastructures, the metadata used to describe the

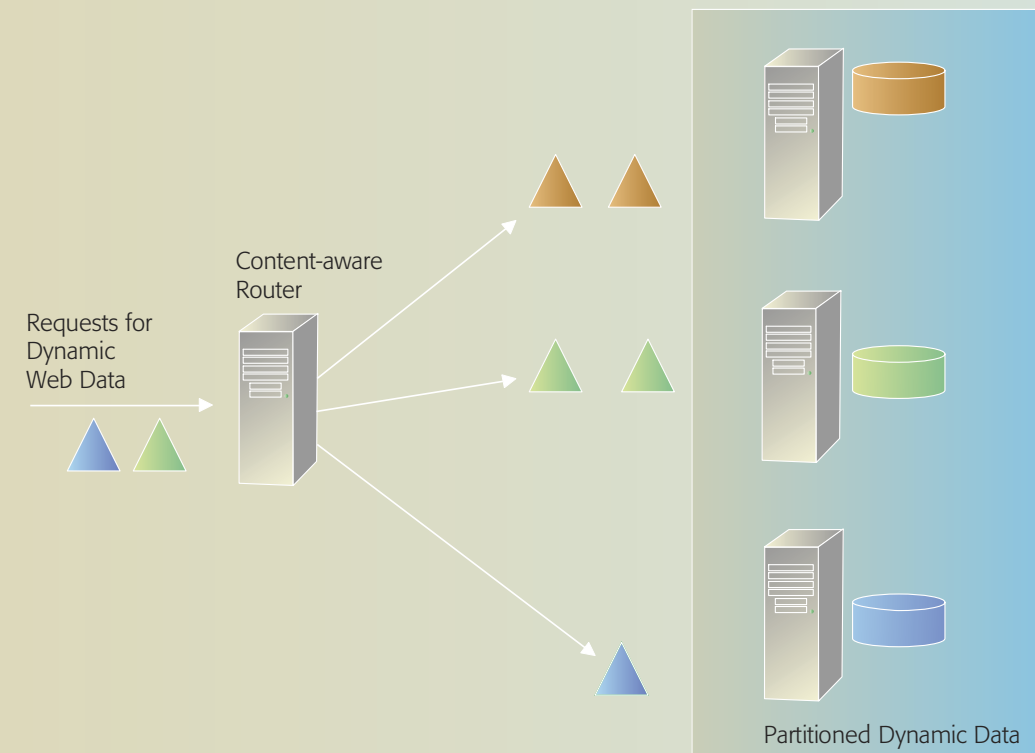
numerous applications and data repositories is likely to use diverse sets of terms. Using a combination of domain-specific and domain-independent ontologies, novel schemes to index and search these metadata repositories are being implemented. In addition, the large overhead of solving information-retrieval problems over large-scale data collections, such as finding categories in a taxonomy spanned by Web-scale search results, are being addressed. Approximate algorithms applied to small random subsets of the search results are being developed, which produce a solution with significantly lower computation, storage, and communication overheads than traditional methods.

PERFORMANCE AND SCALABILITY

Research on improving the performance and scalability of Web applications focuses on multiple aspects. Techniques for efficiently serving dynamic Web data are being developed to improve the performance of Web sites with high amounts of back-end processing by partitioning applications and data across multiple edge and back-end servers. New HTTP-based techniques are also being investigated to increase the energy efficiency of small personal devices that access the Web. An innovative power-aware Web proxy has been devised that can schedule the Web traffic for a device to allow its Wireless Fidelity (Wi-Fi) interface to reduce its energy consumption by switching to lower power states after very short idle intervals; the negative impact of traffic scheduling on latency is compensated by proxy techniques. In the area of Web services, new methods for improving the performance of SOAP-based applications are being studied. New XML template-based techniques have been developed for handling SOAP messages, which result in significant overhead reductions for many interaction scenarios, such as those related to Web services security.



Power consumption profile of a 802.11 client interface when downloading a Web page: benefits of using the Power-Aware Proxy.



Partitioning removes consistency contention and improves throughput