

WebStudio – Building Infrastructure for Web Data Management

Ji-Rong Wen
Microsoft Research Asia
jrwen@microsoft.com

Wei-Ying Ma
Microsoft Research Asia
wyma@microsoft.com

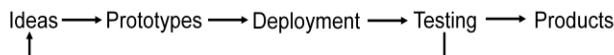
ABSTRACT

To explore various ideas and algorithms for improving relevance of a search engine, we found it necessary to build an infrastructure to provide large-scale data management and data processing capabilities. WebStudio is an infrastructure we have constructed to provide an integrated development environment (IDE) for researchers and developers to use in quickly building prototypes and conducting experiments at Web-scale. It is also a Web data management system to allow users to easily store, access, and manipulate Web data.

1. INTRODUCTION

In this era, a most amazing phenomenon is the Internet's rapid transformation into a computing platform to support programmability and software-driven services on a global scale. Search engine, as a typical Internet service, has become an important infrastructure element for managing Web-scale information and data. Each day witnesses the emergence of many new services, no matter their designation as Web 1.0 or Web 2.0 technologies.

For any Internet company, core competence centers on innovative capabilities and deploying new services quickly to users, along with being able to differentiate companies from competitors. We found that a typical cycle of Web innovation includes the following steps:



That is, the engineers come out with an idea, build a prototype and then deploy it to see how it works, usually with a Beta tag. Users try the service and act as volunteer testers. Soon, engineers receive user feedback, and they use the information to refine the technology iteratively. On the Web, there is a so-called “always beta” phenomenon. The capability to build quick prototypes and conduct experiments is highly important for Internet companies. However, Web innovation usually involves large-scale Web data. The capability of managing, processing and mining such large-scale data is one of the key factors to the success of Internet firms.

It is also well known that doing any experiment at Web-scale is non-trivial. A simple experiment in a small-scale setting usually

becomes impossible when terabytes of data or hundreds of machines are involved. Therefore, we decide to build an infrastructure to facilitate Web innovations. This infrastructure called WebStudio is intended to provide researchers and developers an easy-to-use Web development environment to implement and test ideas at Web-scale. It can be viewed as an integrated development environment (IDE) that provides researchers and developers the needed underlying systems, data, and tools to build prototypes and complete experiments. It is also a Web data management system that enables people to store, access, and manipulate Web data.

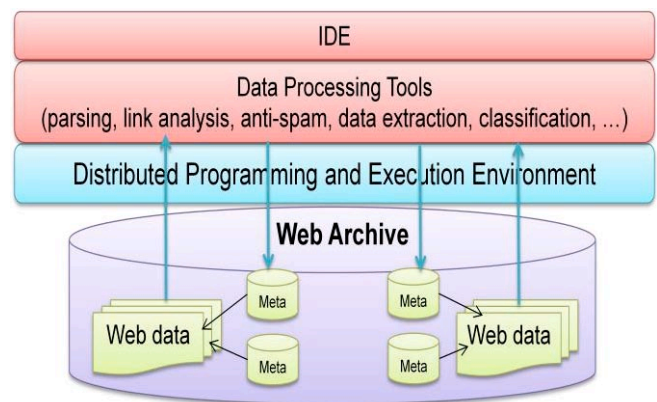


Figure 1. Architecture of WebStudio

2. ARCHITECTURE

Figure 1 shows the overall architecture of WebStudio. It has four layers. First, we have a massive storage system (called Web Archive) for archiving Web data. Upon the storage system, a distributed programming and execution layer is provided to run large-scale data processing tasks. Note that the generated metadata (data derived from data) are stored back to the Web Archive. In addition to the storage layer and distributed environment layer, some common data processing tasks are customized as standard tools, each of which can be shared and reused by others. Finally, we provide a visualized IDE to help users quickly build prototypes and conduct experiments.

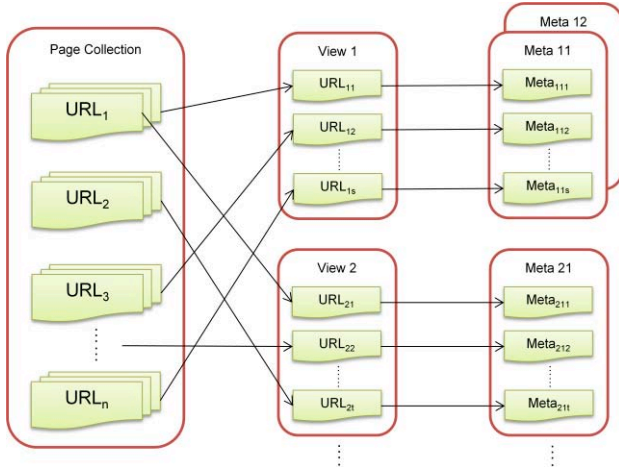


Figure 2. Data Model

3. WEB ARCHIVE

In WebStudio, Web Archive is the place to store and share any persistent data. We archive Web data with arbitrary formats. From these data, different kinds of metadata or knowledge can be deduced by data mining algorithms. Web Archive needs to organize and process both data and metadata. Figure 2 is the typical data model to organize data in Web Archive. Data with different formats and purposes are put into different *collections*. Versioning is supported at the data item level in each collection, and therefore multiple versions of a Web page can be easily managed. Simple SQL-like queries are supported to select data items from collections. Users can create *views* based on query results. We found the “view” concept is a critical factor because people often need to process a subset of data. With “view,” different people can logically work on the same data independently and/or simultaneously. We also provide *metadata collections* to store metadata produced by data processing algorithms. The storing and sharing metadata is important because extracting metadata is expensive when billions of pages are involved. In many cases, we need to reuse metadata for different experiments. The logical relationships among collections, views, and metadata collections are maintained. In this way, the process is simplified for people to locate, reuse, and share data.

4. DISTRIBUTED ENVIRONMENT AND TOOLS

We deploy a powerful distributed programming and execution environment on Web Archive. The distributed environment provides support for people to write distributed data processing algorithms. It allows users to easily read data from Web Archive and write back the produced metadata. Upon the distributed environment layer, Web data processing tasks can be customized as standard tools. The sharing and reusing of tools is also important. For instance, we have developed a powerful HTML

parser to segment a web page into blocks. In WebStudio, this technique is standardized as a tool and it can immediately be used for individual purposes. Otherwise, years could be spent reinventing tools. Tools in WebStudio can be viewed as the counterpart of stored procedures within a database.

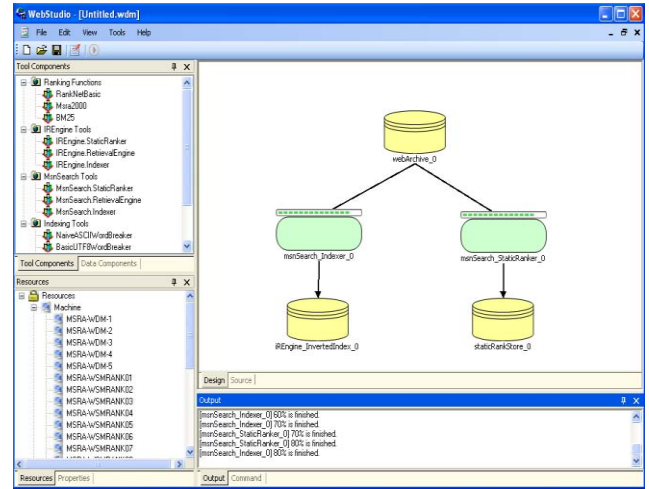


Figure 3. IDE of WebStudio

5. INTEGRATED DEVELOPMENT ENVIRONMENT

To accelerate the prototyping and deployment of Web applications, a visual integrated development environment (IDE) is built for WebStudio, as shown in Figure 3. With IDE, people can use drag-and-drop operations or write tens of lines of code to build a quick prototype by leveraging existing tools and data. Our goal is to build a suite of development tools and a visualization environment in WebStudio that enables users to easily analyze Web data and visualize results using existing functions or user-designed programs.

6. CONCLUSION

To succeed in Web innovation, a large fraction of time could be spent on crawling needed data and building distributed systems to store and manage data while only a small fraction of time is used to do real experiments and attempt to innovate. With WebStudio, we provide the infrastructure for researchers and developers to concentrate on experiments and innovative efforts, and therefore accelerate the pace of Web development. For Web search-related research, this infrastructure facilitates the development and experimentation with new data mining algorithms and allows us to speed innovations to our search engine.

7. ACKNOWLEDGEMENTS

This is a joint work with Shuming Shi, Guomao Xin, Qing Yu, and Yunxiao Ma at Microsoft Research Asia. We would like to thank for their contributions.