From WebArchive to WebDigest: Concept and Examples
(Invited talk)

Li Xiaoming¹  Huang Lianen¹

¹ Institute of Network Computing and Information Systems
School of Electronic Engineering and Computer Science
Peking University
Beijing, 100871 China
Email: {1xm,hle}@pku.edu.cn

Abstract

Much like a black hole, the Web, since its birth, has been absorbing all sorts of data (information) around the globe, ever generated along the path of human civilization. On the other hand, the digitized and networked (webbed) nature of web data, which generally means "easy to access", gives rise to much imagination on re-discovering, re-engineering, and re-using of the oceanic information. Nevertheless, lunch is not free. The same time when we see the grand opportunities, tremendous challenges are ahead. In this talk, I'll first introduce Web InfoMall (http://www.infoMall.cn), the Chinese web archive we have been constructing since 2001. Along with the activities, we observe some useful capabilities have been developed, such as large scale web crawling and very large scale data organization. In addition, we discuss a step beyond the WebArchive, called WebDigest, which is an effort aimed at making use of the data in the web archive. With a web archive and associated capability, "web mining" here has a more or less different meaning, which spans from the structure analysis of the web to named entity and relation extractions, from spatial (if we consider URL as a space) information discovery to temporal information exhibition. The main challenge for us is around the theme of achieving reasonably good performance with affordable cost. As we are from a university lab, the underlying question is: what can be done (and how) in a university lab environment with modest resource. After all, most of the researches started from university lab. We need to understand the feasibilities and compromises while seeing the promises.