Abstract

The exponential expanding of the numbers of web sites and Internet users makes WWW the most important global information resource. From information publishing and electronic commerce to entertainment and social networking, the Web allows an inexpensive and efficient access to the services provided by individuals and institutions. The basic units for distributing these services are the webs sites scattered throughout the world. However, the extreme fragility of web services and contents, the high competence between similar services supplied by different sites, and the wide geographic distributions of the web users drive the urgent requirement from the web managers to track and understand the usage interest of their web customers. This thesis, "**X-tracking the Usage Interest on Web Sites**", aims to fulfill this requirement. "X" stands two meanings: one is that the usage interest differentiates from various web sites, and the other is that usage interest is depicted from multi aspects: internal and external, structural and conceptual, objective and subjective. "Tracking" shows that our concentration is on locating and measuring the differences and changes among usage patterns.

This thesis presents the methodologies on discovering usage interest on three kinds of web sites: the public information portal site, e-learning site that provides kinds of streaming lectures and social site that supplies the public discussions on IT issues. On different sites, we concentrate on different issues related with mining usage interest.

The educational information portal sites were the first implementation scenarios on discovering usage patterns and optimizing the organization of web services. In such cases, the usage pattern is modeled as frequent page sets, navigation paths, navigation structures or graphs. However, a necessary requirement is to rebuild the individual behaviors from usage history. We give a systematic study on how to rebuild individual behaviors. Besides, this thesis shows a new strategy on building content clusters based on pair browsing retrieved from usage logs. The difference between such clusters and the original web structure displays the distance between the destinations from usage side and the expectations from design side. Moreover, we study the problem on tracking the changes of usage patterns in their life cycles. The changes are described from internal side integrating conceptual and structure features, and from external side for the physical features; and described from local side measuring the difference between two time spans, and global side showing the change tendency along the life cycle. A platform, Web-Cares, is developed to discover the usage interest, to measure the difference between usage interest and site expectation and to track the changes of usage patterns.

E-learning site provides the teaching materials such as slides, recorded lecture videos and ex-

ercise sheets. We focus on discovering the learning interest on streaming lectures, such as real medias, mp4 and flash clips. Compared to the information portal site, the usage on streaming lectures encapsulates the variables such as viewing time and actions during learning processes. The learning interest is discovered in the form of answering 6 questions, which covers finding the relations between pieces of lectures and the preference among different forms of lectures. We prefer on detecting the changes of learning interest on the same course from different semesters. The difference on the content and structure between two courses leverage the changes on the learning interest. We give an algorithm on measuring the difference on learning interest integrated with similarity comparison between courses. A search engine, TASK-Moniminer, is created to help the teacher query the learning interest on their streaming lectures on tele-TASK site.

Social site acts as an online community attracting web users to discuss the common topics and share their interesting information. Compared to the public information portal site and e-learning web site, the rich interactions among users and contents bring the wider range of content quality, on the other hand, provide more possibilities to express and model usage interest. We propose a framework on finding and recommending high reputation articles in a social site. We observed that the reputation is classified into global and local categories; the quality of the articles having high reputation is related with the content features. Based on these observations, our framework is implemented firstly by finding the articles having global or local reputation, and secondly clustering articles based on their content relations, and then the articles are selected and recommended from each clusters based on their reputation ranks.

Long Wang
November 3, 2009