Web Design Based on User Browsing Patterns

Yinghui Yang
University of California, Davis, USA

INTRODUCTION

It is hard to organize a website such that pages are located where users expect to find them. Consider a visitor to an e-Commerce website in searching for a scanner. There are two ways he could find information he is looking for. One is to use the search function provided by the website. The other one is to follow the links on the website. This chapter focuses on the second case. Will he click on the link "Electronics" or "Computers" to find the scanner? For the website designer, should the scanner page be put under Electronics, Computers or both? This problem occurs across all kinds of websites, including B2C shops, B2B marketplaces, corporate web-sites and content websites.

Through web usages mining, we can automatically discover pages in a website whose location is different from where users expect to find them. This problem of matching website organization with user expectations is pervasive across most websites. Since web users are heterogeneous, the question is essentially how to design a website so that majority of the users find it easy to navigate. Here, we focus on the problem of browsing within a single domain/web site (search engines are not involved since it's a totally different way of finding information on a web site.) There are numerous reasons why users fail to find the information they are looking for when browse on a web site. Here in this chapter, we focus on the following reason. Users follow links when browsing online. Information scent guides them to select certain links to follow in search for information. If the content is not located where the users expect it to be, the users will fail to find it. How we analyze web navigation data to identify such user browsing patterns and use them to improve web design is an important task.

BACKGROUND

There has been considerable amount of work on mining web logs – Web Usage Mining. Web usage mining is a viable framework for extracting useful access pattern information from massive amounts of web log data for the purpose of web site personalization and organization (Missaoui et al 2007; Srivastava 2000; Nasraoui et al. 2003; Mobasher & Anand 2005). Various tasks can be achieved via web usage mining (e.g., finding frequent and interesting navigation patterns, predicting future page requests and page recommendations). Spiliopoulou et al. (1998) and Spiliopoulou et al. (1999) propose a "web utilization miner" (WUM) to find interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the human expert using WUM's mining language. Chen & Cook (2007) proposes a new data structure for mining contiguous sequential patterns from web log. Liu et al (2007) presents a study of the automatic classification of web user navigation patterns and propose an approach to classifying user navigation patterns and predicting users' future requests. The approach is based on the combined mining of web server logs and the contents of the retrieved web pages.

A small subset of the research in web usage mining uses the usage patterns observed from the web logs to improve web design. There is quite some research on link recommendations based on users' previous browsing patterns, mainly utilizing the sequence of pages visited. Perkowitz et al. (1998) and Perkowitz et al. (1999) investigate the problem of index page synthesis, which is the automatic creation of pages that facilitate a user's navigation of a website. By analyzing the web log, their cluster mining algorithm finds collections of pages that tend to co-occur in visits and puts them under one topic. They then generate index pages consisting of links to pages pertaining to a particular topic. Baraglia & Silvestri (2007) introduces a web personalization system that is online and incremental and it is designed to dynamically generate personalized contents of potential interest for users of large web sites made up of pages dynamically generated. It is based on an incremental personalization procedure tightly coupled with the web server. It is able to update incrementally and automatically the knowledge base

based on the ontology underlying the site. Data mining techniques can be applied to enriched web logs to extract knowledge that could be used to improve the navigational structure as well as exploited in recommender systems.

Another stream of related research is on Information foraging theory (Pirolli & Card 1995; Pirolli & Fu 2003). As mentioned in Juvina & Herder (2005), web sites generally are designed for a general audience with varying goals. As it is hard to satisfy all categories of users with one design, adaptive hypermedia systems try to better support the users by personalizing content or link structure. Traditional techniques in the latter category involve link hiding, sorting, annotation, direct guidance and hypertext map adaptation (Brusilovsky 2001). When trying to find information related to a task, users have to rely on proximal cues such as the link anchor text to decide what their next action will be. If the proximal cues are not clear enough, or if the users do not have sufficient insight on the structure of the site, they may become disoriented, i.e. they don't know their current position in a web site, how they came to that point or where to go next (Herder & Juvina 2004). Various studies have been carried out to infer user goals from their actions (e.g. Chi et al. 2003). Given these goals, the utility of the various navigation options on a web page can be estimated (Kitajima et al. 2000; Pirolli & Fu 2003) and communicated to the user by means of link relevancy indicators, or link suggestions.

Also, research on users' page revisit behavior has documented that users frequently visit pages already visited before. Earlier studies (Catledge & Pitkow 1995; Tauscher & Greenberg 1997; Cockburn & McKenzie 2001) have shown that the majority of page requests involve requests to pages visited before.) Tauscher & Greenberg (1997) identified the following main reasons for revisiting pages: (a) the information contained by them changes; (b) they wish to explore the page further; (c) the page has a special purpose; (d) they are authoring a page; (e) the page is on a path to another revisited page. When the revisited pages are home pages and index pages that serve to navigate users to a number of pages, the reason for backtracking can be also to search for information that's not found at the current location (Srikant & Yang 2001).

There is a great amount of research on locating documents via searching (Yeung et al 2007, Heflin et al 2003). Searching eventually becomes the dominating way of finding information when the web structure is too deep and complex to browse. Another stream of related research is automatic acquisition of taxonomies or concept hierarchies from a text corpus (Cimiano et al 2005). This can help organize documents automatically in a structure.

MAIN FOCUS OF THE CHAPTER

Our focus in this chapter is to address the gap between the web-site designer's expectations and the web users' expectations. The web designers' expectations are observed from the web site structure, and web users' expectations are observed from their usage patterns discovered for web logs. Through web usages mining, we can automatically discover users' traversal paths from which we can infer their expectation about how the web site should be structure.

We first briefly discuss several papers that address this problem and then focus more on one particular paper to illustrate a simple idea that can be used to improve the web design through users' behavioral patterns.

Nakayama et al. (2000) tries to discover the gap between the web-site designer's expectations and user behavior. Their approach uses the inter-page conceptual relevance to estimate the former, and the inter-page access co-occurrence to estimate the latter. They focus on website design improvement by using multiple regressions to predict hyperlink traversal frequency from page layout features. Paik et al (2002) describes the design and the implementation of a system through which existing on-line product catalogs can be integrated, and the resulting integrated catalogs can be continuously adapted and personalized within a dynamic environment. They propose a methodology for adaptation of integrated catalogs based on the observation of customers' interaction patterns. Gupta et al (2007) proposes a heuristic scheme based on simulated annealing that makes use of the aggregate user preference data to re-link the pages to improve navigability. Organizations maintain informational websites. The information content of such websites tends to change slowly with time, so a steady pattern of usage is soon established. User preferences, both at the individual and at the aggregate level, can then be gauged from user access log files. Their scheme is also applicable to the initial design of websites for wireless devices. Using the aggregate user preference data obtained from a parallel wired website, and given an upper bound on