ABSTRACT

Researchers spend considerable time searching for relevant papers on the topic in which they are currently interested. Often, despite having similar interests, researchers in the same lab do not find it convenient to share results of bibliographic searches and thus conduct independent time-consuming searches. Research paper recommender systems can help the researcher avoid such time-consuming searches by allowing each researcher to automatically take advantage of previous searches performed by others in the lab. Existing recommender systems were developed for commercial domains to assist users by focusing towards products of their interests. Unlike those domains, the research paper domain has relatively few users when compared with the huge number of research papers. In this paper we present a novel system to recommend relevant research papers to a user based on the user’s recent querying and browsing habits. The core of the system is a scalable subspace clustering algorithm, SCuBA (Subspace ClUstering Based Analysis) that performs well on the sparse, high-dimensional data collected in this domain. Both synthetic and benchmark datasets are used to evaluate the recommendation system and to demonstrate that it performs better than the traditional collaborative filtering approaches when recommending research papers.
BACKGROUND AND MOTIVATION

The explosive growth of the world-wide Web and the emerging popularity of e-commerce has caused the collection of data to outpace the analysis necessary to extract useful information. Recommender systems were developed to help close the gap between information collection and analysis by filtering all of the available information to present what is most valuable to the user (Resnick and Varian, 1997).

One area of the Web that has seen continued growth is the online publication of research papers. The number of research papers published continues to increase, and new technology has allowed many older papers to be rapidly digitized. A typical researcher must sift through a large quantity of articles manually, relying on keyword-based searches or paper citations to guide him. The search results of researchers with similar interests can help direct a more effective search, but the process of sharing search results is often too cumbersome and time consuming to be feasible. A recommender system can help by augmenting existing search engines by recommending papers based on the preferences of other researchers with similar interests. We assume that the proposed system is localized in a typical lab setting which augments an existing search engine. Rather than performing search in some digital library, SCuBA tries to take advantage of the existing search results. It tries to find similar interest groups of users based on their browsing patterns and recommends research papers which might be interesting to them.

There are two main branches of recommender systems; content based filtering and collaborative filtering. Content based filtering (CBF) approaches create relationships between items by analyzing inherent characteristics of the items. Collaborative filtering (CF) systems do not analyze an items properties, but instead take advantage of information about users’ habits to recommend potentially interesting items. The analysis of user behavior patterns, allows collaborative filtering systems to consider characteristics that would be very difficult for content based systems to determine such as the reputation of the author, conference, or journal. CF approaches are also well suited to handle semantic heterogeneity, when different research fields use the same word to mean different things.

In many domains, there is an ever increasing number of users while number of items remains relatively stable. However, in domains such as research paper recommendation (Singh et al., 2005), the number of users (researchers) is much less than the number of items (articles). Collaborative filtering systems face two major challenges in the research paper domain: scalability to high dimensional data and data sparsity. In a typical recommender system there are many items. For example, Amazon.com recommends specific books of interest from a large library of available books. Item-based approaches that determine similarity measures between items do not perform well since the item space is extremely large. A user based approach allows us to leverage the relatively small number of users. An intuitive solution used by early collaborative filtering algorithms is to find users with similar preferences to the current user and recommend other items that group of users rated highly. Even with a relatively small number users, however, this approach is computationally complex. The use of clustering algorithms to pre-determine groups of similar users has been used to significantly increase performance (Ungar and Foster, 1998; Mobasher et al., 2000).

A particular user of the system will probably purchase a very small percentage of the available books. As a result, if we consider the data as a user-item matrix, a typical row will be extremely sparse, with only a few columns containing 1’s, representing purchased books. Similarly, in the domain of research papers where a researcher will be interested in articles related only to a particular research area, the row representing the researcher will also be very sparse. Notice that the number of
Related Content

A New Framework for Intelligent Semantic Web Services Based on GAIVAs

Adaptability and Adaptivity in The Generation of Web Applications

Class Level Test Case Generation in Object Oriented Software Testing

Improving Privacy and Security in Multicloud Architectures