User Centered Interactive Search and Browsing

Yi Zhang, University of California Santa Cruz

1 Executive summary

Traditional Information Retrieval (IR) systems such as Google and Yahoo have gained wide popularity by letting users access information using a query. However, a user often does not issue a query when he has an information need, especially when he does not know exactly what needs to be known. Besides, even given the same query and information goal, a recent study shows that the relevance judgements varies greatly among different users.

On the other hand, the main information access mechanism within an enterprize is traditional file directory browsing instead of keyword search. As users interact with file systems of ever increasing size, it is becoming more difficult for them to familiarize themselves with the entire structure and content of the system.

The proposed project will tackle this challenge by focusing on developing intelligent interactive search and browsing techniques to help users find the information they are looking for from billions of non-relevant files. To achieve this goal, we will research a search portal that can provide: 1) meta data based navigation and search (faceted search); 2) personalized ranking of documents/files; and 3) automatic structured query recommendation. Personalized and collaborative search are the main themes of the proposed research portal. We will model the preferences a user based on the past history of the user. Other users's search history will further improve the search quality of a user. This will enable the system automatically guess the intention of a user and thus provide an intelligent personalized context sensitive interactive navigation/search interface to help the user find information needed. The project could greatly improve people's ability to access information in a petabyte-scale systems.

With ISSDM grant, the PI is planning to collaborate with LANL Technical staff. One LANL employee who are planning to pursue Ph.D. degree through the joint ISSDM has contacted the PI for advising opportunity. If there is a good match, the PI will create education and supervising opportunities for LANL staff. If funded. the PI and the graduate student sponsored have the intention to visit LANL. The PI are wiling to teach LANL-simulcast courses.

2 Project description

How to effectively and efficiently find and integrate needed information that is strewn across heterogeneous and constantly changing information resources distributed across a diverse group of division within an organization is a challenging problem. A typical employee within an enterprise keeps a list of intranet web pages, file servers and directories, and visits them as needed to find required information. This approach is ad-hoc and time-consuming and risks missing important information. An individual's list of resources may be incomplete, outdated, or too long to search often, especially when one is busy.

The proposed research will remove some of the limitations that make it inconvenient or difficult for employee to find information from distributed information resources by building a portal that provides a user with a single access point to distributed information resources. Through this portal, users interactively search and browse their shared petabyte-scale file system.

To develop this portal, we will research the following three topics in the following years.

2.1 Faceted search

Faceted search interfaces are becoming a popular method to allow users to interactively search and navigate complex information spaces. These interfaces present users with key-value metadata to be used in query refinement. For example, the customer can narrow down the list of candidate products by putting constraints over the category facet, price facet, brand facet, and age facet at www.toysrus.com.

Within the information retrieval community, faceted search has been primarily focused on static document collections, such as libraries and e-commerce catalogs, not with the dynamic collections such as those found in shared storage. In the file system community, work on semantic file systems has primarily been focused on low-level issues such as load balancing and system APIs, not on the end-user application level. While popular in e-commerce and digital libraries with limited facet types, not much research has been conducted on how to use this information seeking mechanism in peta scale file system with hundreds and thousands of potential facet types. This proposed work intends to bridge this gap.

In the proposed project, we will advance the faceted search research and create “smart” interactive feedback mechanism. We will divide the search user interface into two separate parts. The first part is a list of documents ($D_{recommended}$) that the system believes are the most likely to relevant. The second part is a list of facet-value pairs $V_{recommended}$ that the system believes could aid the user in finding the target documents. (The size of $D_{recommended}$ and $V_{recommended}$ are tunable parameters that the interface designer will specify). A user can take the following actions to interact with the system: 1) Provide relevance judgements for some documents in $D_{recommended}$; 2) Provide relevant judgements for some facet-value pairs; or 3) Stop.

We model the whole process as a Bayesian decision process. Getting the feedback on facet-value pairs or recommended documents reduces the system's uncertainty of whether a document matches the profile and increases the expected utility. The challenge is not to overwhelm the user with hundreds of possible facet-value pairs while asking for the feedback.

We will tackle this challenge through personalization. In a shared petabyte-scale system, different groups of users will have differing needs. By customizing the the search/browse interface to each user, the user can have a detailed view of the portion of the file system he/she is most concerned with, without becoming overwhelmed by extraneous portions of the file system.

Personalization is most effective when the system has large amounts of user feedback in order to learn a model for each user. This presents a two problems. First, users must often endure a period of poor performance before the user sees any benefit from personalization. The second problem is given the size of the file collection, each user will only interact with a tiny fraction of the available files. We propose to solve these problems by using a combination of content-based and collaborative recommendations. Content-based recommendations measure the similarity of the internal structure, in this case the faceted metadata, of relevant files to make suggestions to a user. Collaborative recommendations on the other hand, make suggestions based on the similarity of the queries (or user groups) and which files were considered the most relevant to each query. This hybrid approach has found success in large scale e-commerce systems for document recommendation, and can be adapted to the faceted search domain.

2.2 Learning to provide personalized ranking of results

Proper ranking of results for queries for file system is open problem in the IR community. Enterprise search is even harder than traditional web search, because that there is few information about the relationships among the stored files. To provide a good ranking function in an enterprise environment, we will research adaptively learning to provide user specific ranking.

The core of a retrieval system is the ranking function that order documents according to their degrees of relevance, preference, or importance as defined in a specific context. One major approach use to find the ranking function is to automatically learn a function from training data. However, the criteria of importance, relevance, preferences are different for different users within an enterprise.

The propose project will research how to learn user specific ranking function while observing and interacting with the user. A probabilistic ranking model will be developed, and this model will have the following desirable characteristics:

Using the content of a document The text and meta data of a document will be used to measure how relevant a document with respect to a user query.

Using user features User features, such as the location of the user and the affiliation of the user, will be used when they are available

Collaborative filtering (learning from others) It may take a while before the system can gather enough data from the user and learn a reliable user specific ranking function. However, a good initial performance is the incentive for a new user to continue using the system. The ranking function of a new user will be learned using information from others.

Using social networks The social context of a user is often very helpful for inferring the preferences of this user.

Using heterogeneous feedback from the user The system may collect other feedback from that user. For example, the user may explicitly rate the quality of a result, provide implicit feedback by clicking a document, or reading a document for a while. These feedback will be integrated to together to learn the user specific ranking function

2.3 Structured query recommendation

We will explore how to improve existing search engines with proactive structured query recommendation. Instead of waiting for the user to create structured queries through advanced search interface manually, the search engine can let the user progressively narrow down or modify the search query by choosing one out of several automatically recommended structured queries.

In our preliminary research, we have developed a personalized search engine that does semi structured query recommendations and tailers the expanded query to the user's information needs based on the information on the user's local PC. The search engine recommends the expanded query as facet constraints to the user and lets the users take control of personalization. For example, when a user inputs a query "Britney Spears", the system recommends "Britney Spears format:video" to limit the search results to video clips if most of the user's local documents related to "Britney Spears" are video clips. In a pilot user study, the approach has statistically significant improvement over Google and Yahoo.

In the proposed project, we will go further and explore techniques to create "smart" search engines that can take the initiative to recommend structured queries to the user actively. The system will be able to recommend structured queries that contain document content, document facets, resource facets, and context facets. Getting the value of each facet reduces the system's uncertainty of whether a document matches the information needs of a user and increases the expected utility. The major challenge is not to overwhelm the user with hundreds of possible facets.

We will adapt the Bayesian active learning approach the PI proposed to the facet recommendation problem and trade off exploration and exploitation.

The proposed research is expected to bring significant benefits to the average user. Because structured query expansion is not a well studied problem in the IR community and has several advantages: 1) it solves millions of search engine users' immediate information need; 2) it may help a user better understand how the engine works through the advanced queries; and 3) it may train users in how to form better queries over time.

2.4 Evaluation

We will evaluate interactive search engine using standard information revaluation measures such as precision and recall. We will also use a new measure we proposed to quantify the amount of effort required by a user to satisfy his/her information need. The new measure is a utility defined over the actions a user performs over the course of a search session. A user's information need can be expressed as a subset of the documents in the corpus being searched, and the number of documents of this subset that must be retrieved by the user. Given this assumption, the success of a user in meeting his/her information need can be measured as the number of relevant documents successfully retrieved, divided by the number relevant documents required to be retrieved. Since the number of documents required for a completely successful search can exceed one, and that faceted search interaction is essentially a series of query refinements, evaluation must be with respect to a search session, instead of an individual query.

For each user search session, assume user takes a sequence of T actions $(a_1, a_2, ..., a_T)$. At each time point t, the user takes an action a_t, which changes the state of the user from q_t to q_{t+1}. We define the user utility for this session as:

$$U = \sum_{t=0}^{T} a_t R(q_{t+1}, a_t, q_t) \tag{1}$$

q_t is the state of the interface at time t. We represent q_t using a combination of the current query, the system suggested queries, and the system suggested documents at time t. $reward q_{t+1} a q_t$ represents the reward that the system receives when the user transitions from state q_t to q_{t+1} via action a. For example, the reward could be -1 if the user click a link and find no relevant document. If the user click a link and find a relevant document returned, the reward will be positive.

2.5 project status and accomplishments

This proposed research is multi year project. It is based on the project previously funded by ISSDM summer 2007. The project has lead to the following two refereed research papers.

PDSW07/SC07 Jonathan Koren, Yi Zhang, Sasha Ames, Carlos Maltzahn, Ethan Miller. Searching and Navigating Petabyte Scale File Systems Based on Facets. In Proceeding of 2007 ACM Petascale Data Storage Workshop, International Conference for High Performance Computing, Networking, Storage and Analysis.

This paper introduces faceted search and outlines three current research directions in adapting faceted search techniques to petabyte-scale file systems: first, how to create faceted meta data, second, how to index meta data, and third how to build intelligent interface for user to explore the data space based on facets.

WWW08 J. Koren, Y. Zhang, and X. Liu. Personalized Faceted Search. To appear in Proceedings of the 17th International Conference on World Wide Web.

While faceted search is popular in e-commerce and digital libraries, not much research has been conducted on which metadata to present to a user in order to improve the search experience and how to evaluate a faceted search interface. We propose using collaborative filtering and personalization to customize the search interface to each user's behavior. We also propose a utility based framework to evaluate the faceted interface. A set of four algorithms for generating the faceted search interface are proposed and evaluated using the novel evaluation methodology.