Improving Web Search Results using Clickthrough history

U. Rohini

May 4, 2007

There has been a tremendous growth in the amount of information on the WWW. Information retrieval systems are critical for overcoming this information overload and providing the information of interest to the users of the systems. Although many information retrieval systems (e.g., web search engines and digital library systems) have been successfully deployed, the current retrieval systems are far from optimal. A major deficiency of existing retrieval systems is that they generally lack user modeling and are not adaptive to individual users. This inherent non-optimality is seen clearly in the following two cases: (1) Different users may use exactly the same query (e.g., Java) to search for different information (e.g., the Java island in Indonesia or the Java programming language), but existing IR systems return the same results for these users. Without considering the actual user, it is impossible to know which sense Java refers to in a query. (2) When a query contains partial information: When a query contains an acronym or a short cut, there is not sufficient information required to infer the information need of the user. Existing IR systems return mixture of results containing the exact word which might contain different expansions. For example a query like "SBH" can mean "State Bank of Hyderabad" or "Syracuse Behavioral Healthcare" among others. Knowledge of the interest of the user can be helpful in acquiring more knowledge and hence return better results to the user. Thus using user context information about the user and the query is necessary for improving the retrieval performance. Indeed, personalized search essentially boils down to capturing and exploiting related user context information of a query to improve search accuracy.

While there are some existing work in the personalized search, it is far from a solved problem. In this thesis, we studied the problem of personalized search. The problem was found to be much complicated than it appeared to be. The major challenges for personalized search were identified to be two fold. First is, modeling the appropriate context of the user and learn a user model from them. Second is, how to utilize the user model to improve the search accuracy. Another important challenge is the evaluation of the experiments. There are no standard and bench mark datasets available on which the experiments can be peformed. This makes comparison with earlier work in the literature and replicating their results difficult. There are also no standard metrics available to effectively evaluate personalized search algorithms. Commonly used metrics used to evaluate Information Retrieval systems are usually borrowed and used.

We propose three methods for personalizing of search results. Each of the three methods, has two phases. The first phase called *User Profile Learning* or *User Modeling*, we learn a model of the user representing his potential interests from the past search history. The first two methods are based on utilizing the past search history consisting of the past queries and their corresponding clickthroughs. The third method attempts personalization based on this past queries alone. The second phase is called *Reranking* we use the user model learnt is used to personalize the search results.

- Personalized Search using language modeling techniques
  We propose two algorithms using statistical language modeling techniques. The first is using simple n gram techniques where we compute a language model which probabilistic distribution of words from the user training data. In second algorithm based on language modeling we view the process of query generation through a noisy channel model.

- Personalized Search using Machine learning Methods
  The second method is based on machine learning algorithms. We propose an approach to personalized based on Ranking SVMs, a variation of the Support Vector Machine learning algorithm which can learn from the partial feedback data present in the clickthrough data of users as mentioned above for improving web search. We used ranking SVMs to learn user profile from the clickthrough history of the user. Rereanking is done using the weight vector obtained from training the ranking SVM. We experimented with different features and feature weights for learning the model.

- Personalized Search without Relevance Feedback
  The third method is another interesting method where we model the user based on just the past queries posed by the user. We do not use the implicit feedback provided by the user we rather only used the past queries by the user. The simple approach worked suprising very well proving that queries alone contain a potential and rich source of user context information. It poses an intersting question: Can Personalization of Search be done without Relevance Feedback ?

We performed experiments on data collected over 3 months by a popular search engine. We performed our experiments on 17 users from the query log. The first two months of data is used to learn a profile and the third month data is used for evaluation. Our experiments showed that our approaches outperformed the baseline with a wide margin

The contributions from this thesis are as follows. A suit of algorithms with the following: 1) Basic IR aproaches 2) proposed and baseline personalization approaches. In addition, the tool kit also has 3) Evaluation metrics for IR and personalized search and 4) Synthetic data creation. Also several observations were made from several studies during this thesis which provide interesting directions to personalized search research. The two important such observations are Can Personalization of Search be done without Relevance Feedback ? and is Can Simulation of User Search Behaviour be done.