

Extracting expertise to facilitate exploratory search and information discovery: Combining information retrieval techniques with a computational cognitive model

Wai-Tat Fu & Wei Dong
Applied Cognitive Science Lab
University of Illinois at Urbana-Champaign
405 N. Mathews Avenue, Urbana, IL 61801
wfu/wdong@illinois.edu

ABSTRACT

We compared and combined the traditional information retrieval (IR) methods of expert identification with a computational cognitive model to test their effectiveness in facilitating exploratory search performance using a data set from a large-scale social tagging system. We found that the two methods of expert identification, although based on different assumptions, were in general consistent in extracting useful structures for exploratory search. The methods, however, did show systematic differences in their effectiveness to guide users to find popular vs less popular topics. The findings have important implications on presentations of information cues that facilitate interactive IR and discovery.

Categories and Subject Descriptors

H.5.3 Group and Organization Interfaces: Collaborative computing. H5.4. Information interfaces and presentation (e.g., HCI): Hypertext/Hypermedia.

General Terms

Performance, Design, Human Factors, Theory

Keywords

SNIF-ACT, Knowledge exploration, knowledge exchange, social tagging, Expert Identification

1. INTRODUCTION

The Web has become a participatory social-computational systems that allow people to explore, learn, and share information with others. A good example is social bookmarking systems such as del.icio.us, CiteULike.org, and Bibsonomy.org, which allow users to annotate, organize and share their web-based resources using short textual labels called tags. Many have argued that social tagging systems can provide navigational cues or “way-finders” to facilitate exploratory search and information discovery [8-9]. The notion is that, given that social tags are labels that users create to represent topics extracted from Web documents, interpretation of these tags should allow other users to better predict and evaluate relevance of documents in interactive search.

Many researchers have argued that the openness of social tagging systems may result in a large number of low-quality tags that are not meaningful to other users. Although many methods have been proposed to distinguish between indices and contents contributed by experts and novices, there is still a lack of systematic evaluation on how the extracted expertise could be utilized to facilitate knowledge exploration by others. The goal of this position paper is to demonstrate how a computational cognitive model of Web search could be utilized to complement existing data-mining techniques to predict the usefulness of different

presentations of expert-generated indices and contents extracted by different methods in facilitating exploratory search.

Research has shown that the definition of expertise can be *referential* or *representational*. In the referential definition, experts are individuals who are recognized and referred to by others. The idea is that the more a person is being referred to, the more likely that others will follow and regard the person as an expert. Many information retrieval (IR) methods rely on this definition of expertise. Typically, the referential method identifies experts based on the hubness or authoritativeness of the source by analyzing the link structures in the system [10]. In contrast, in the area of HCI or cognitive sciences, definition of expertise is often representational [3]: experts tend to show better search performance or have better domain knowledge. Methods based on the representational definition of expertise use certain forms of semantic representations to extract structures and indices in Web resources and to identify users who share similar semantic representations [1, 7]. This method, compared to the referential method, has the advantage of being able to measure how well users can *interpret* and *represent* different topics by tags in a social tagging system, but may not capture as much the “social” aspect of the definition of expertise as in IR methods.

2. The simulations

2.1 The Database

We used the database dump provided by Bibsonomy on January 1st of 2009, which contains 3859 users, 201,189 tags, 543,43 resources, and are connected by 1,483,767 tag assignments. We selected the most recent 6 months of tag assignments in our simulations, which contained data from 537 users, 18,278 tags, 52,098 resources, and connected by 101,428 tag assignments.

2.2 Expert identification by Link Structures

We chose to identify experts using the SPEAR algorithm [10], which used mutual reinforcement to generate the lists of experts and quality resources in the folksonomies. Following Noll et al. [10], the lists were represented as two vectors: E represented the vector of expertise scores of users, i.e., $E = (e_1; e_2; \dots; e_M)$, and Q represented the vector of quality resources, i.e., $Q = (q_1; q_2; \dots; q_N)$, where M and N were the total number of users and resources in the set respectively. Mutual reinforcement was implemented by preparing an adjacency matrix A of size $M \times N$, where $A_{ij} = 1 + k$ if user i had assigned a tag to document j , and k users had assigned tags to document j after user i , and $A_{ij} = 0$ otherwise. Thus, if user i was the first to tag resource j , A_{ij} would be set to the total number of users who tag resource j ; but if user i was the last one, then A_{ij} would be set to 1. This effect of this was to create a bias to those users who discovered quality resources. Following Noll et al., in order to balance the impact of the

discovery and hubness effect, the value of A_{ij} was adjusted by the square root function, such that $A_{ij}' = \sqrt{A_{ij}}$. Based on this adjacency matrix, the calculations of expertise and quality scores followed an iterative process similar to that of the HITS algorithm. However, because the SPEAR algorithm also took into account the time of tagging as a factor that influenced expert identification, it was less susceptible to spammers (who typically give a lot of tags to a wide range of resources, but are less likely to be the first to identify quality resources). Specifically, in each iteration, E and Q were updated as:

$$E' = Q \times AT \text{ and } Q' = E \times A$$

The final lists of E and Q would represent the expertise and quality scores of the users and resources, which could be sorted to identify the top experts and resources in the system.

2.3 Identifying semantic structures

To study the differences in the semantics structures of the resources tagged by experts and non-experts, we extracted topics from the resources using the LDA model [1]. However, because topic extraction is computationally extensive, we selected the top 5,000 quality resources identified by the SPEAR algorithm, and then randomly sampled another 5,000 resources from the resources located in the bottom half of the quality vector Q. We called these the *high-quality* and *low-quality* sets of resources. In addition, we identified the first 50 experts from the SPEAR algorithm, and then randomly sampled another 50 users in the bottom half of the expert vector E. We extracted the resources tagged by these experts and non-experts to form the *expert* and *non-expert* sets of resources. We then processed the HTML files based on the URLs of the resources in the database. We filtered out any non-English pages and pages that contained fewer than 50 words, and eventually obtained 5000 usable resources from each of the four sets. We performed the topic extraction algorithm derived from the standard LDA model on each set of resources.

We were interested in how tags given by experts and non-experts and those on low and high quality resources could serve as good navigational cues for the users. To measure this, we assumed that users would adopt a tag-based topic inference process [5-6], which allowed them to predict whether the tagged resource would contain topics that they were interested in. This value could be calculated by the posterior probability $p(c_j|tags)$, where c_j is the topic of interest. For the current purpose, it was useful to compare the predictive distribution of tags in each set of resources to compare the usefulness of the tags. This empirical distribution $P(\text{tag}_i|c_j)$ could be derived from the LDA model, but due to space limitation the exact derivation is not given here. The predictive distribution of tags in each topic will show whether there are differences in the predictive power of tags in each set. The assumption is that the higher the predictive power, the more likely users would be able to use the assigned tags to infer what topics can be found in the resources, and thus the more useful will be the tags in guiding users to find relevant information in the system. In other words, the predictive probability of tags reflected the quality of tags for knowledge exploration

Another useful measure for understanding the semantic structures in the different sets of resources is to compare the predictive distributions of *topics* in the different set of resources. This probability $P(c_j)$ can also be estimated empirically based on the LDA model. The predictive distributions of topics reflected how likely certain topics could be found in a resource. Comparing the predictive distributions of topics between the set of resources

would therefore show how the distributions of popular or "hot" topics would correlate with the experts and quality resources identified by the SPEAR algorithm.

2.4 Simulating search by computational cognitive models

To simulate exploratory search, we first randomly pick a topic (represented as a topic-word distribution) and a random tag, and calculate average performance of the model-searcher in different search environments through repeated simulations. The model-searchers were developed based on previous research, and due to space limitation we could not repeat the details here [4-5], but they were developed based on the cognitive mechanisms that were shown to match well with actual users as they performed Web search. The model-searcher would navigate in the folksonomies and collect resources that were relevant to the topic of interest. Topical relevance of a resource was calculated by Kullback-Leibler (KL) divergence between the desired and the best matching topic-word distribution in the resource. If the KL divergence reached the threshold, the resource was considered relevant. To measure exploratory search performance, we limited each model-searcher to perform 1000 steps of "clicking" to count the number of relevant resources it could find. In other words, we assumed that the average number of relevant resources found within 1000 transitions in the hypergraph reflected how well the environment could support exploratory search conducted by the model-searcher. We also randomly selected the topic of interest based on the predictive probabilities of topics, such that half of the simulations were looking for popular topics, the other half were looking for less popular topics. We could then compare how different model-searchers would perform differently when searching for popular or less popular topics.

We also created a number of search environments by ranking different navigational cues based on the predictive probabilities of the tags, experts, and resources. For example, in a "tag" environment, all tags were ranked according to $p(c_j)$ calculated by the LDA model, in a resource/experts environment, resources/experts were ranked by the quality/experts scores calculated by the SPEAR algorithm. Simulations results would therefore allow comparisons of how these cues generated by the IR techniques and the LDA model were *utilized by real users* to demonstrate their usefulness in facilitating exploratory search.

3. Results

3.1 Topics distributions

Figure 1 shows the mean predictive probabilities of topics (top) and tags (bottom) against the ranked list of resources in the expert and non-expert sets of resources (left) and against the ranked list of experts in the high- and low-quality sets of resources (right). These probability distributions show that the semantic structures in each set of the resources identified by the referential method had very different properties. In terms of the predictive probabilities of topics, low-ranking resources found by experts tended to contain more "popular" topics than non-experts, but this difference seemed to diminish quickly as the resource quality rank increased (top-left of Figure 1). On the other hand, comparing the sets of high and low quality resources identified by the referential method, there seemed to be consistently more popular topics in the high quality resources than low equality resources, and this difference seemed to be relatively insensitive to the rank of experts within the sets (top-right of Figure 1).

The predictive probabilities of tags also showed interesting difference between the sets of resources. In general, resources tagged by experts contained more predictive tags compared to those by non-experts, and this difference was relatively stable across the set of resources (bottom-left of Figure 1). Similarly, the predictive probabilities of tags between the high and low quality sets of resources also showed differences, but this difference diminished quickly as the expert rank increased.

To summarize, results implied that while following the list of high quality resources would allow users to discover mostly "hot" topics, following the list of experts would allow users to sometimes discover "cold" topics (but could be useful for a subgroup of users) in the folksonomies

3.2 Tag Distributions

Figure 1 (bottom) also shows that while tags created by experts were in general predictive, not all tags in the high quality resources were predictive. To confirm these differences, we compared the empirical probability distribution functions (PDF) of the predictive probabilities of topics and tags in each sets of resources (see Figure 2). One could see that the topic distributions between experts and non-experts were indeed less distinguishable than those between low and high quality resources (top-left of Figure 2), but the reverse was true for the tag distributions (bottom-left of Figure 2). This suggested that quality of resources were in general better at predicting "hot" topics; but high quality resources did not necessarily contain fewer "hot" topics. Rather, expert-generated tags tended to be more predictive of "cold" topics than resource quality.

It was quite possible that, because the SPEAR algorithm (and the referential method in general) primarily determined the quality of a resource based on the number of times the resource was identified by experts, resources that contained popular topics were more likely reckoned high quality. As popularity of topics in a folksonomy tended to correlate with frequencies of occurrence of corresponding real-world events, ranking of resources based on the referential method would therefore likely benefit users who were interested in following "hot" topics. On the other hand, although experts identified by the referential method were users who frequently tag the high quality resources, they were also users who tend to tag many resources, and many of these resources were not tagged by other experts. We therefore see that rankings by resources tended to be generally better at distinguishing "hot" from "cold" topics than by the distinction between experts and non-experts.

3.3 Exploratory search performance

Figure 3 shows the number of relevant resources found by three model-searchers in the exploratory search simulations. The position-searcher always followed the ranking of cues during evaluation, the topic-satisficer combined ranking and topical relevance of the cues during evaluation (see [4] for details), and the perfect-searcher always picked the cue that had the highest topical relevance (with respect to the topic that it was searching). Comparisons of the three model-searcher would therefore reveal the effects of position and topical relevance of cues on search performance.

As shown in Figure 3, the general uptrend from all four figures suggested that the addition of more rankings of tags, experts, and resources had led to better exploratory performance. Interestingly, for popular topics, rankings of users and resources seemed to lead to slightly better results than the ranking of tags. Consistent with

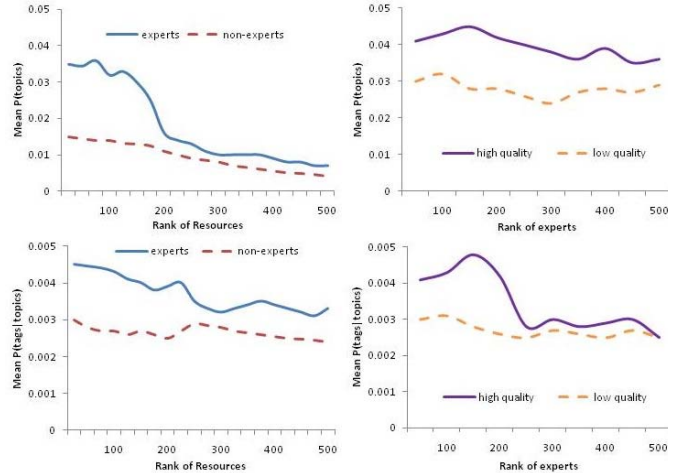


Figure 1. Predictive distributions of topics (top) and tags (bottom) plotted against ranks of resources and experts (lower rank is better) in the sets of resources tagged by experts and non-experts (left) as identified by the algorithm, and the sets of resources identified as high and low quality by the algorithm (right).

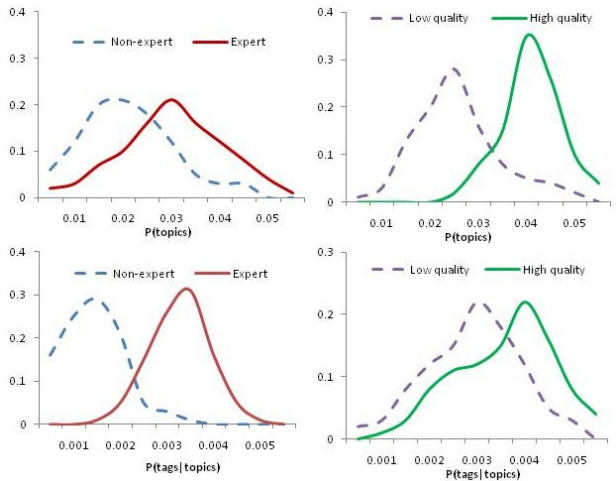


Figure 2. The empirical PDF for the predictive probabilities of topics and tags in each of the four sets of resources.

previous results, this could be attributed to the fact that the rankings of users and resources were based on the referential method that were in general better at predicting hot topics. In contrast, ranking of tags was based on their predictability of topics, which depended, to a large extent, on the likelihood that the tags were uniquely associated with the different topics. Given that hot (popular) topics tended to be associated with semantically general tags that appeared in multiple topics [2], the general predictability of tags for popular topics were therefore lower than the rankings derived from link structures.

Compared to the position-searcher, the topic-satisficer in general found more relevant resources, suggesting that the process of sequential topic evaluation improved exploratory search performance. Similar to the position-searcher, however, tag ranking was slightly less useful for exploring for hot topics compared to expert and resource rankings. On the other hand, the combination of tag and expert or tag and resource rankings did significantly improve performance. Performance of the topic-satisficer in the all-ranking environment was almost the same as

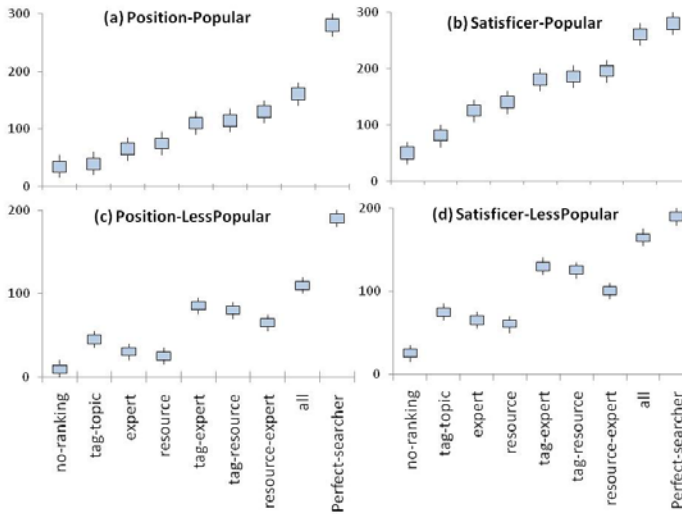


Figure 3. Exploratory search performance of the position-searcher (left) and the topic-satisficer (right) when the topics of interest were popular (top) and not popular. (bottom).

the perfect-searcher, which always picked the most predictive nodes in every transition.

When the model-searchers explored for less popular topics, the results were similar but did show differences (bottom of Figure 3) when compared to exploration for hot topics. In particular, while the addition of rankings improved performance, but for both model-searchers, *tag ranking* in general led to better exploratory performance than expert and resource rankings. This could be attributed to the fact that for cold topics, the predictive power of tags was higher than that by expert and resource rankings (see top two graphs of Figure 1). Following the tags therefore led to a higher chance of discovering cold topics than by following expert and resource rankings.

In summary, the patterns of simulation results showed not only that expertise rankings could improve exploratory performance, but it also showed that different expert identification methods could have systematic differences in their influence on exploratory performance. In particular, we found that while expert and resource rankings based on the referential method could facilitate exploration of hot topics, ranking of tags based on the probabilistic topic extraction method could facilitate discovery of cold topics. In our simulations, a combination of the two methods seemed to lead to the best overall result in providing effective navigational cues that facilitate knowledge exploration. Future research should focus on how to adapt the presentation of these cues based on interaction patterns of the user to allow the user to select different cues for exploration of cold or hot (or both) topics.

4. Conclusion and General Discussion

The current results provide support to the promising aspect of using expertise in social tags to facilitate exploratory search. Specifically, our results showed that (1) the method based on the referential definition of expertise was more useful for generating rankings that facilitate search for popular topics, while the method based on the representational definition of expertise was more

useful for generating rankings that facilitate search for less popular topics, (2) rankings of tags based on their predictive probabilities of topics could facilitate search of less popular topics, and (3) combinations of referential and representational methods of expert identification could facilitate knowledge exploration of both popular and less popular topics.

We have shown how IR methods can be combined with semantic and mechanistic models of exploratory search to understand how different interface representations could impact overall utilities of the system. Our results highlight the importance of including realistic assumption of the users to evaluate the functional utilities of structures extracted from different data-mining methods. In general, we believe that it is useful to investigate the dynamics between how individual users will actually *utilize* information cues or structures extracted from a social information system, and how they would in turn influence the computational properties of the system itself.

5. REFERENCES

- [1] Blei, D. *et al.*, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
- [2] Ericsson, K. A., and J. Smith, eds., "Toward a general theory of expertise: Prospects and limits," Cambridge, MA: Cambridge University Press, 1991
- [3] Fu, W.-T. *et al.*, "A Semantic Imitation Model of Social Tag Choices.," *Proceedings of the IEEE conference on Social Computing*, pp. 66-72, Vancouver, BC, 2009.
- [4] Fu, W.-T., and P. Pirolli, "SNIF-ACT: A cognitive model of user navigation on the World Wide Web," *Human-Computer Interaction*, vol. 22, pp. 355-412, 2007.
- [5] Fu, W.-T. *et al.*, "Semantic Imitation in Social Tagging," *ACM Transactions on Computer-Human Interaction*, in press.
- [6] Kang, R. *et al.*, "Exploiting Knowledge-in-the-head and Knowledge-in-the-social-web: Effects of Domain Expertise on Exploratory Search in Individual and Social Search Environments," *Proceedings of the ACM Conference on Computer-Human Interaction*, Atlanta, GA, 2010.
- [7] Landauer, T. K., and S. T. Dumais, "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, pp. 211-240, 1997.
- [8] Marchionini, G., "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41-46, 2006.
- [9] Millen, D. R. *et al.*, "Social bookmarking and exploratory search."
- [10] Noll, M. G. *et al.*, "Telling experts from spammers: Expertise ranking in folksonomies," *Proceedings of the ACM conference on Information Retrieval (SIGIR)*, pp. 612-619, MA: Boston, 2009.
- [11] Russell, D. M. *et al.*, "The cost structure of sensemaking," in *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, Amsterdam, The Netherlands, 1993