Explicit Graphical Relevance Feedback for Scholarly Information Retrieval

Shaoshing Lee, Indiana University Bloomington
Chun Guo, Indiana University Bloomington
Xiaozhong Liu, Indiana University Bloomington

Abstract
In this paper, we present a new method to collect users' feedback on scientific heterogeneous graph to enhance the scientific information retrieval performance. Meanwhile, a new search system is implemented to validate the new feedback hypothesis. Unlike earlier approaches, by using the new search system scholars can mark the useful/not useful venues, papers, authors, and keywords on a heterogeneous graph, and the feedback algorithm can select the optimized paths on the graph to enhance the retrieval performance.
Keywords: information retrieval; feedback; graph mining; interface
Citation: Lee, S., Guo, C., Liu, X. (2015). Explicit Graphical Relevance Feedback for Scholarly Information Retrieval. In *iConference 2015 Proceedings*.

Contact: li415@indiana.edu, chunguo@indiana.edu, liu237@indiana.edu

1 Introduction

In the past few decades, the volume of scholarly publications has increased dramatically, which has had a significant effect on how scholars perceive, retrieve, and consume publications. However, characterizing high-quality scientific information needs (given a textual query) can be more complex and challenging than for other domains. Sometimes, textual queries cannot adequately represent what scholars are looking for, especially when (junior) researchers venture into unexplored academic realms where they are ill prepared.

Recently, some studies (Liu, Guo, Yu, & Sun, 2014) have shown that heterogeneous bibliographic networks can be constructed by utilizing multiple types of links from the scientific repository. It has been demonstrated that by using the heterogeneous link information in networks, mining functions, such as similarity search, ranking, clustering, and classification can be significantly enhanced. However, to the best of our knowledge, few prior studies have addressed the "ad-hoc" academic search or recommendation problem from a relevance feedback (RF) perspective. How to utilize the heterogeneous graph-based RF is not trivial.

Take the text query "relevance feedback with language model" as an example. Classical search or feedback algorithms are able to find the papers similar to the query. Feedback based on heterogeneous graph, however, provides a different result set, i.e., "*Latent Concept Expansion Using Markov Random Fields*", which comes from the complex relations and paths on the graph and not necessarily similar to the initial query. From a RF's viewpoint, given the target query, we can first retrieve a number of top-ranked papers, and then locate important/unimportant papers, authors, citations, topics, and venues (nodes) on the graph via the paths to those nodes. For instance, given the aforementioned query plus graph-based RF, the system can find "*Latent Concept Expansion Using Markov Random Fields*", because it is related to "*Croft, W. B*" (an important author node given the query), and "SIGIR" (an important venue node). In other words, the heterogeneous graph-based RF conceptualizes various kinds of paths on the graph as the ranking functions (or features) and different paths can "vote" for the recommended citations through a learning model. Comparing with text-based search and RF, heterogeneous graph-based RF offers more "global scholarly information". Experiments (Liu et al 2014) show that heterogeneous graph-based PRF (pseudo RF) outperforms text-based PRF by 36.2% for mean average precision (MAP) and 15.9% for (normalized discounted cumulative gain (NDCG) for citation recommendation.

In this study, we propose a new method and a new scholarly search system to enable users to provide various kinds of graphical (explicit) relevance feedback information, and enhance the search experience and performance.

2 Literature Review

Relevance feedback is an effective re-ranking method to improve the retrieval performance. However, earlier experiments also show that text-based relevance feedback approaches, i.e., Rocchio's query

expansion and term reweighting method (Rocchio, 1971), do not perform well or even harm the ranking performance in some search scenarios due to the noisy top-ranked documents. Collins-Thompson et al. (2009), for example, used multiple sources of domain knowledge or evidence to enhance the robustness of pseudo feedback by characterizing feedback gain, feedback benefit and feedback risk while minimizing uncertainty in the dataset via risk-aware algorithms.

Graph-based feedback or pseudo feedback is a new ranking assumption (Liu et al., 2014), which is rooted in topology-based search. For instance, Dean and Henzinger (1999) found the disadvantage of user formulated queries that can hardly characterize information needs, and they utilized connectivity between web pages to recommend related websites to users on the basis of their initial URLs. However, graph-based feedback is not well studied in previous researches. Vassilvitskii (2006) investigated the new feedback method by employing hyperlink based web-graph distance for relevance feedback in web searches. The experiment results showed that, for web search, graph-based feedback outperforms standard text-based relevance feedback methods. Unlike web search, academic retrieval and citation recommendation provide a more complex search scenario, in that, the candidate papers can be interlinked in a heterogeneous graph.

3 Methodology and System Design

The main functionality of the novel system is to present the search result based on user's keyword and enable user to provide relevance feedback. The system will search and present four types of data: *paper, author, venue*, and *keyword*, and collect user's explicit feedback. The design of the system focuses on 1) visualizing the data on a heterogeneous graph to assist the users to explore the result and 2) capturing the users' relevance feedbacks of the result. The first goal is highly important and closely related to the second goal.

The novel search interface, as Figure 1 shows, visualizes structural data and graphical data. Structural data allows easy browsing of the search result, and graphical data depicts the relations between different kinds of metadata. The two data sections also work as a reference to each other.

Figure 1 shows the interface when user input a query. It is split into two areas: 1) a list of result items representing their hierarchical structure (Figure 1, 1), and 2) a visual graph representing the relations of the result items (Figure 1, 2). There are four types of items in the result list: paper, author, venue, and keyword; each of them has a different color indicated at the bottom right of the interface (Figure 1, 3). By hovering over any node (shown as circle) in the graph (Figure 1, 4), the selected node will be enlarged, and its related node will be highlighted. Item in the list is also highlighted to help the user find out the context of the selected node. Same thing applies to the items in the list: hovering over any item (Figure 1, 5) will highlight the node in the graph.

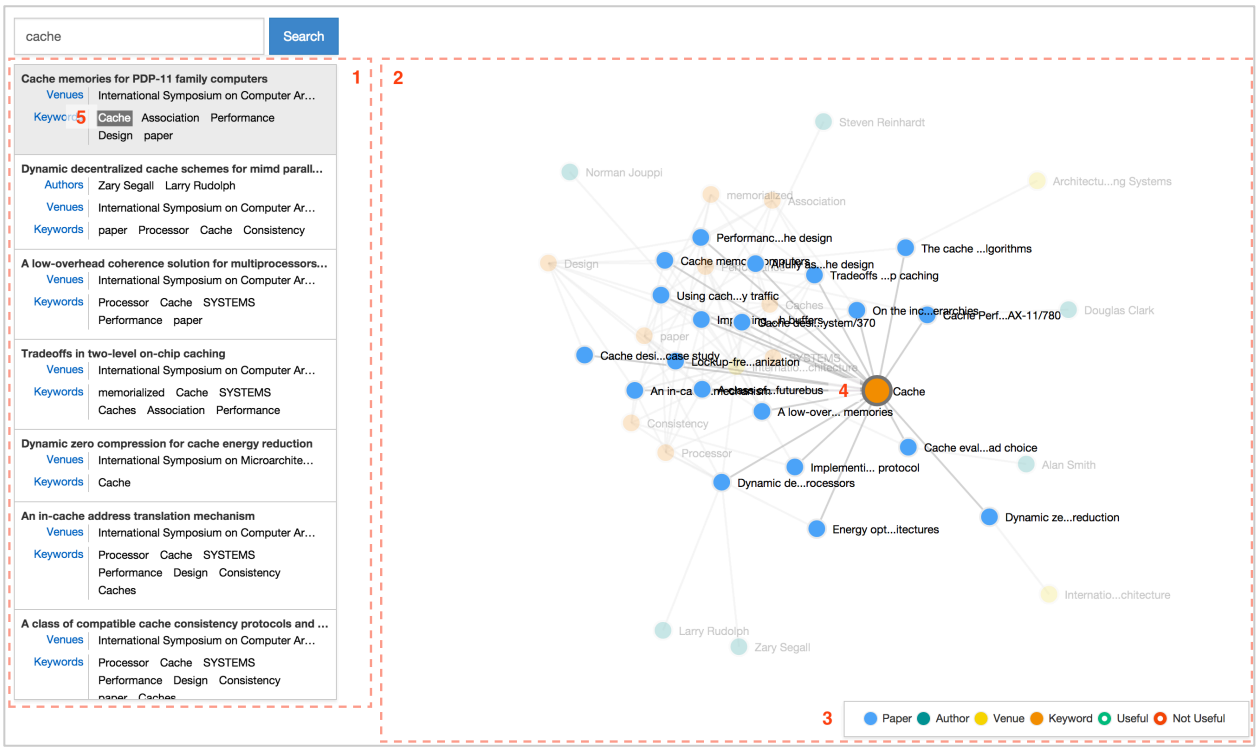

Figure 1. Search result visualization

3.1 Relevance Feedback

By clicking on an item from the list or the graph, an inspector panel will show up (Figure 2, 1) with the selected item and its related items. Each item in the inspector panel has a checkbox on the left, and user can select items to mark as "**useful**" or "**not useful**" by clicking on one of the two buttons below (Figure 2, 2).

The items that are marked as "useful" will be circled by a green stroke (Figure 3, 1), and the "not useful" items will be circled by red strokes. All the feedbacks provided using the inspector panel are pending unless the user confirms the changes by clicking the buttons beside the search input box (Figure 3, 2). This gives the user a second chance to review his changes or undo any mistakes.

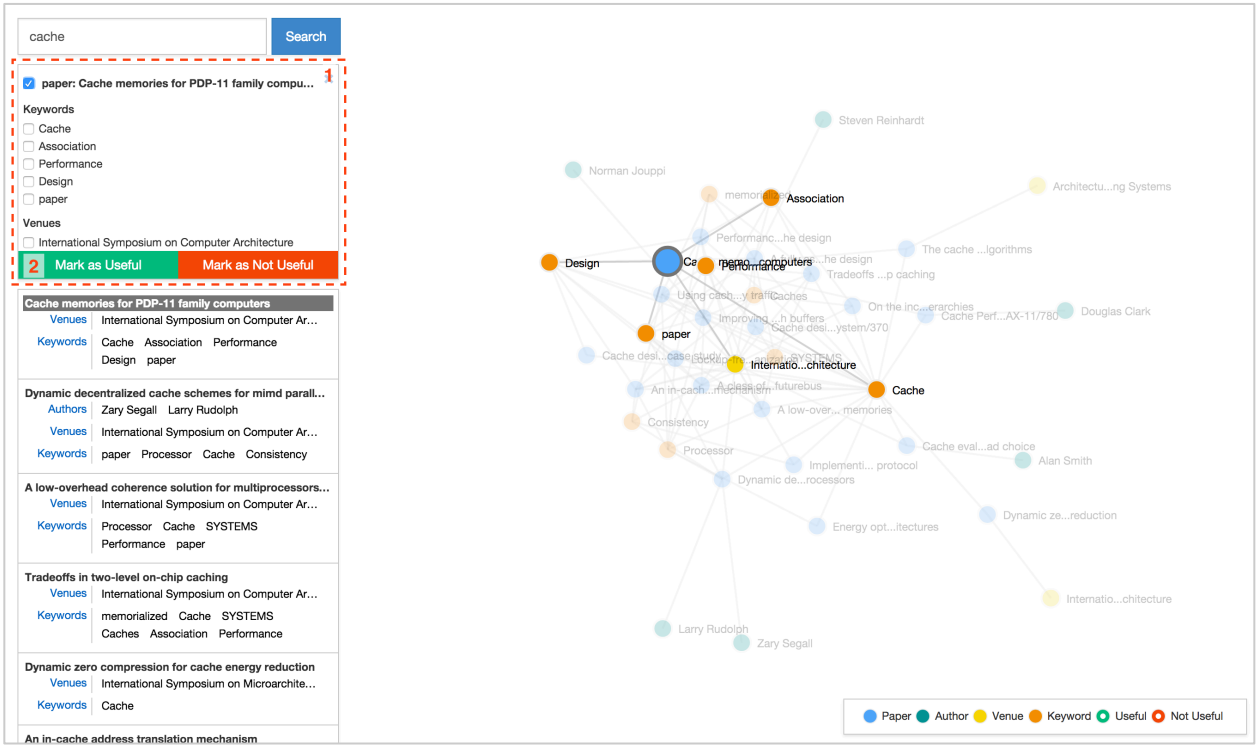

Figure 2. User feedback collection

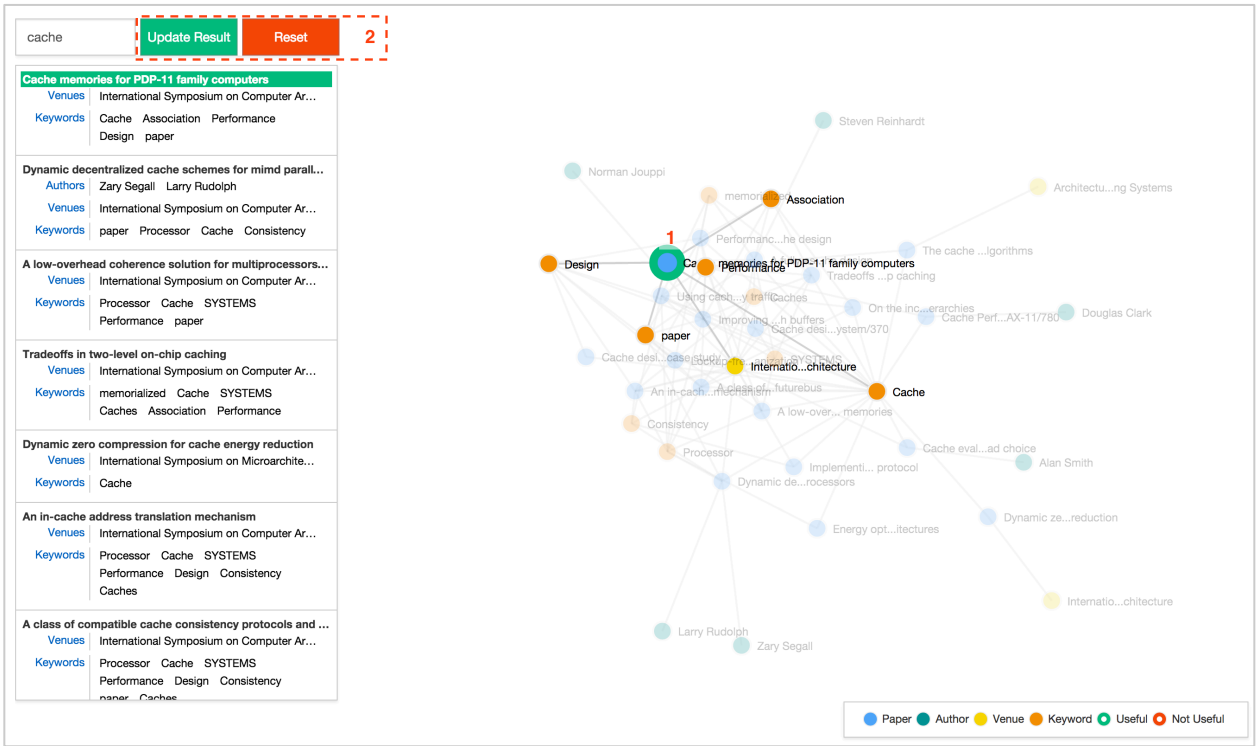

Figure 3. Update the search result

3.2 Feedback Model

From heterogeneous mining viewpoint, we propose a number of path-based ranking features based on Liu et al. (2014). Basically, the newly added papers have a high probability to random walk to the original relevant papers. By using this interface, we can further enhance this model that all the newly retrieved papers should highly likely walk to the useful nodes, while not likely walk to the not useful nodes.

$$P(p|q,F) = \max\{\sum_{i=1}^{|F_{pos}|} P(p \rightsquigarrow F_i) - \sum_{j=1}^{|F_{neg}|} P(p \rightsquigarrow F_j)\}$$

, where the probability that a paper p is relevant the given query q and feedback F equals the maximum likelihood that the node p random walk to all the positive (useful) nodes minus the random walk probability to the negative (not useful) nodes on the graph.

4 Conclusion

In this study, we propose a new search and feedback method by using heterogeneous graph mining. Meanwhile, we develop a new system to enable users to efficiently provide feedback information on the visualized graphical search result. Unlike earlier approach, user feedback is no longer restricted to text information. Instead, different kinds of nodes on the heterogeneous will tell the importance of different paths, which can be used to calculate the random walk probability from the candidate paper nodes to all the (positive and negative) feedback nodes.

References

Collins-Thompson, K. K. (2009). Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th International Conference on Information and Knowledge Management*, (pp. 837-846). doi:10.1145/1645953.1646059

Dean, J., & Henzinger, M. R. (1999). Finding related pages in the World Wide Web. *Computer Networks*, *31*(11-16), 1467.

Liu, X., Guo, C., Yu, Y., & Sun, Y. (2014). Meta-Path-Based Ranking with Pseudo Relevance Feedback on Heterogeneous Graph for Citation Recommendation. In *Proceedings of the 23rd International Conference on Information and Knowledge Management*. (pp. 121-130). Shanghai, China.

Rocchio, J. J. (1971). Relevance Feedback in Information Retrieval. In G. Salton (Ed.), *The SMART Retrieval System - Experiments in Automatic Document Processing*: Prentice Hall.

Vassilvitskii, S., & Brill, E. (2006). Using web-graph distance for relevance feedback in web search. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 147-153). Seattle, Washington, USA.

Table of Figures