A graph based approach of solving cold start problem

by

Xu Shen

a dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science
Department of Computer Science
New York University

May, 2014

Abstract

Traditional keyword-based Search has been popularized by commercial search engines like Google and Bing, where retrieval is heavily dependent on keywords and their frequency in target documents. However, such techniques often fail to capture the informational needs of the users, resulting in failure to retrieve essential information even when it is available. Faceted search enables users to navigate a heterogeneous information space by combining text search with a progressive narrowing of choices along multiple dimensions. However, they suffer from the cold start problem of a non-domain knowledge searcher could not provide any effective keywords for the initial searching. In this paper, we propose a new approach to overcome the latter problem by leveraging facets as an implicit feedback for detection of critical facets to identify the best search results. We present a prototype implementation of faceted search and show how traditional faceted search can be augmented by use of implicit feedback.

Contents

Listing of figures

This is the dedication.

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

0

Introduction

Information seekers trying to gauge public opinion or learn about current events face a torrent of information from tens of millions of sources worldwide. Many existing search engines equate user information needs with a keyword query, assuming that a user knows what words to use to best describe his or her information need. However, many common information retrieval tasks do not fit into this traditional keyword search paradigm. Some information needs are not naturally representable as queries. In other cases, an information need may have a natural query, but is too complex to be expressed as keywords.

Today, keyword-based search engine is arranged manually - suitable keywords or keyword chains are assigned to information resources by trained experts -, sophisticated algorithms are used to generate keywords automatically form (textual) information resources, e.g., researchers have proposed some approaches trying to use user profiles for search engine to provide personalized search result[1,14]. However, assigning appropriate keywords remains expert knowledge, i.e. the ordinary user hardly knows anything about the keywords, which are require to actually find a specific resource. Even worse, the user can never be sure about the completeness and the integrity of the achieved search results.

Part of the responsibility for that situation bears the traditional keyword-based search paradigm. You have to know the appropriate keywords to find a specific resource. That's all. But, not all search engine users have the same information needs, because users might have different ways to search for information. Moreover, if the user tries to achieve an overview of actually available information about a certain topic, today's web search engines are flooding search results by millions. Traditional keyword-based search does not consider user's higher level abstraction desire and result ranking is mainly based on link popularity.

Faceted search has gained great success in exploration domain over the past years, and most popular online urban guide websites, such as Yelp, now provides faceted search interfaces. On faceted-search-enabled websites, users can narrow down the list of interesting places by putting constraints on a group of merchandize facets, such as categories, services, reviews, products, etc. Well designed faceted search has been shown to be understood by the average user[5]. However, faceted search is also used as auxiliary search tool for the keyword search. This is also known as cold start problem which involve a degree of automated data modeling. Specifically, it concerns the issue that the system cannot draw any inferences for users or items about which it has not gathered sufficient information.

This motivates us to explore whether we can adapt the faceted search idea to the general purpose document retrieval. Users might have preferences for certain document facets. For example, online buyers might have preferences on brands, colors, etc. In all these cases, users have clear ideas about some facets of their interested documents, and this information might help the system learn users' preferences and interests. Ideally, users would provide structured queries to describe their information needs more accurately. However, when a user browses the web at different times, he or she could be accessing pages that pertain to different topics. Different categorical data cannot represent a different purpose for a user. However, different kinds of interests might be motivated by the same kind of interest at a higher abstraction level. That is, a user might possess interests at different abstraction levels, and the higher-level interests are more general, while the lower-level ones are more specific.

In this paper, we explore a implicit interactive user feedback mechanism based on facets to solve the cold start problem. In this mechanism, instead of letting users provide relevance feedback on documents or create structured queries actively, the system models the documents as a weighted graph, where the weights are the similarity scores based on users interests. To achieve this goal, the system suggests the faceted constraints (in the form of facet-value pairs) and user can choose interesting facet-value pairs to improve the returned documents.

The proposed faceted feedback mechanism may have the following advantages. First, the suggested facet-value pairs are usually short and easy to understand. Compared with document-based feedback, this may reduce the cognitive overload of the user and thus is more likely to be adopted by the average user. Users can quickly select multiple facet-value pairs in a short time, so the system might get more user feedback. Second, it may help a user better understand the corpus, how the engine works, and train users in how to form better queries.

The rest of this paper is organized as follows. In section 2, we talk about the related work. Section 3 is the focus of this paper, and describes the faceted feedback mechanism. We propose four facet-value pair recommendation methods and two retrieval models in this section. In section 4, we describe the methodology of our experiments. Section 5 gives the experimental results and the corresponding analysis. Section 6 concludes this paper.

1

Foundations and Related Work

This section introduces faceted search technology and the prerequisites to implement *Only-Facets Search*. Furthermore, the concept of exploratory search and implicit feedback mechanism are explained.

1.1 Exploratory Search and Faceted Search

In the contrast to traditional keyword-based search, exploratory search assists the user in exploring the data space to improve search experience. Thereby, the user is able to navigate the search space as well as to reorganize the content and user interface for her own needs with appropriate interactive elements. Exploratory search is mainly used for searching to learn or to investigate, which involves multiple iterations and return sets of objects that require cognitive processing and interpretation[8]. To implement explorative search, the underlying data needs to be full made accessible. One way to establish a exploratory search is to reorganized and to filter the search results according to these relationships by so-called faceted search.

For example, Hearst et al. developed flamenco, a multi-column faceted spatial browser for hierarchical faceted metadata[4]. Petratos described facets as conceptual categories, which are created to organize the presentation of all available data into an easy to view concise set of conceptual groups[9]. Furthermore, faceted search also means to discover new association and new kinds of knowledge.

There are many open research questions about how to generate useful groups and how to design interfaces to support exploration using grouping. Currently, faceted search is quite popular. The representation know as faceted metadata is gaining great traction within the information architecture and enterprise search community[6].

1.2 Cold Start Problem in Faceted Search

Recommender systems suffer from the cold start problem of a new user who start with an empty profile and encounters a difficulty of communication with his community members. Many approaches have been proposed[K.W. Leung & Chung,13] as the approach of Z. Zaier[15] who has studied the challenges of recommender systems namely the cold start. In our work, we leverage the graph method for exploring related entities based on the user's interest score. However, there are little researches based on graphs for the cold start problem in faceted search area.

According to the same perspective and to solve the problem of cold start document, Roy et al. have proposed another approach based on minimum effort drill-down[10]. Roy proposed to ask searchers to make relevance judgements about returned objects and then executing a revised query based on the judgement. This solution provides a dialog with the user to extract more information from her on other desired attribute values. This approach depends on user interaction which requires strong human participation in a more continuous and exploratory process. However practice shows that people are often unwilling to take the added step to provide feedbacks when the search paradigm is the classic turn-taking model. Our developed approach differs from this prior work along several key dimensions: (a) our proposed approach considers implicit feedback based on users' choice, (b) our proposed approach is graph based and depends on user interaction, (c) our algorithms can work in conjunction with available ranking functions.

1.3 Implicit Search Feedback

Relevance search feedback is a commonly used query refinement technique that can be traced back to 1960s. The basic idea is to reply on user interactions to better capture the user information need. Document-based relevance feedback is one of the most widely used explicit feedback mechanisms. Many approaches have been proposed to incorporate document relevance feedback into retrieval[16,17]. Our work is motivated by early work in relevance feedback, and differs by focusing on implicitly retrieving user interest score with faceted metadata.

2

Higher Level Behavior Modeling

This chapter is intended to deal with the process of modeling user's higher abstraction level behavior and its implementation. To begin with, an introductory example is presented and the functionality of the prototype is explained.

2.1 Relationships of Lower Level and Higher Level Behavior

User's interests change over time, studies of user search behavior have a long history in Information and Library Science[2 D.E 12]. When a user browses the web at different times, she could be accessing pages that pertain to different topics. For example, a user might be looking for research papers at one time and airfare information for conference travel at another. That is, a user can exhibit different kinds of interests at different times, which provides different contexts underlying a user's behavior. However, different kinds of interests might be motivated by the same kind of interest at a higher abstraction level. That is, a user might posse interests at different abstraction level - the higher level interests are more general, while the lower-level ones are more specific.

During a browse session, general interests are in the back of one's mind, while specific interests are the current foci. In this paper, we focus on implicit methods for incremental creating an ordered representation of user profiles. Utilizing an interest score, has been proven to be successful for the evolution of personal interest[11].

2.2 Hits Queue Model for Facets

In the traditional faceted search, each category is represented by a facet, user can choose different facets to drill down the search results, however, different facets directly present user's interests. As a motivate example, let us consider a scenario that an user wants to find a restaurant where is near by a famous place she might want to visit after lunch. To find the most suitable restaurant, she needs to first search for a restaurant then clicks on different facets to filter the searched result. Multiple clicks present her interest score, however unless she could construct a complicated query, she could not easily represent her interest.

This common scenario inspires us to map user's clicks into a *Hits Queue*. The underlying data structure is a priority queue. A priority queue is a queue for which each element has an associated priority, and for which the dequeue operation always removes the lowest (or highest) priority item remaining in the queue. In this paper, each element in the queue presents for a chosen facet with interest score.

An chosen *facet* is presented by a entity in the priority queue. An *priority* contains a set of attribute-value paris <M, V>, where ai is an attribute name and Vi is a set of values v, v2, ..., vk. Each value vj belongs to Vi is either an atomic value.

2.3 Mapping clicks to Hits Queue

To map <M, V> pairs, the hits queue is generated with the interaction processing while the user clicking on different facets. As user dynamically change their preferences during the time, the hits queue . Furthermore, the number of hits, when clicking for the entity, is stored as freq(md). We construct the queue from users clicks as follows:

- From each click c on a facets f, add
-

3

The Search Results Modeling

3.1 Data Graph Modeling

Let *D be a set of search results. D can either be a base relation or a materalized view or it can be the result of query Q. A node in data graph D is assigned to a weight to determine its importance in the dataset. While in the hits queue, multiple facets stores user's interest score which determine the importance of the entity types, in the data graph they capture the* *relative importance among entities of the same type.*

We model entities and references using a weighted undirected graph. A data graph D is a weighted undirected graph in which we repent:

- *Each entity of the search results by a node.*
- *Each relationship between two nodes by a edge.*
- *Each confidence score between two nodes by a weight (weight of the edge which links them).*

Note that the edges weights are modeling interest relationships of similarity between the individual tastes which are not constant. In fact, these weights express the mutual trust between paris of actors. We chose to restrict the values of these indices between (-1 and 1), where ”1” is a very strong link between two users (positive relationship) and ”-1” is a negative relationship.

3.2 Exploration Model

We determine the edge weights in D using the interests score in the corresponding hits queue Q. We assign edge weight in the data graph using the weight of the most interested score among nodes in Q.

4

Conclusion and futrue work

In this work, we have addressed the problem of how to improve faceted search for navigational and exploratory search by using implicit feedback mechanism and demonstrated an improved exploratory search with an evaluation of the search process. We have show how to use graph based mode to enable a simple faceted search. By using this, we were able to make implicitly existing relations among multiple data sets explicit and to augment the ordinary *keyword-based search by presenting additional related information and resources to the user via an appropriate interactive user interface.*

Faceted search is at it's early stages as a research data. Currently, there does not exist an overall accepted best-practice neither on how to realize nor on how to evaluate. Although, we have obviously increased the recall of obtained results by providing an faceted search interface, the precision of the suggested resources has to be determined by the user and her personal information needs.

Improvements of the graphical user interface explicitly supporting the investigative and navigational aspect of our approach will be considered in future work. For better support in data space navigation, future work is focussed on the combination of faceted and explorative search features to satisfy the searchers curiosity and to foster serendipitous discovery.

Overall, we have implemented a first prototype for exploratory faceted search, which gives the user the possibility to discover resources that are usually hidden aways from the user's eyes in the search engine index.

A

Some extra stuff

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum *bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.*

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis.

Cras sed ante. Phasellus in massa. Curabitur dolor eros, gravida et, hendrerit ac, cursus non, massa. Aliquam lorem. In hac habitasse platea dictumst. Cras eu mauris. Quisque lacus. Donec ipsum. Nullam vitae sem at nunc pharetra ultricies. Vivamus elit eros, ullamcor-

per a, adipiscing sit amet, porttitor ut, nibh. Maecenas adipiscing mollis massa. Nunc ut dui eget nulla venenatis aliquet. Sed luctus posuere justo. Cras vehicula varius turpis. Vivamus eros metus, tristique sit amet, molestie dignissim, malesuada et, urna.

Cras dictum. Maecenas ut turpis. In vitae erat ac orci dignissim eleifend. Nunc quis justo. Sed vel ipsum in purus tincidunt pharetra. Sed pulvinar, felis id consectetuer malesuada, enim nisl mattis elit, a facilisis tortor nibh quis leo. Sed augue lacus, pretium vitae, molestie eget, rhoncus quis, elit. Donec in augue. Fusce orci wisi, ornare id, mollis vel, lacinia vel, massa.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim *wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.*

References

[1] *B. Tan, X. S. & Zhai, C. (2006). Mining long-term search history to improve search accuracy.* In Proceedings of KDD, *(pp. 718–723).*

[2] *Bates, M. (1979). Information search tactics.* Journal of the American Society for Information Science, *(pp. 205–214).*

[D.E] *D.E, R. Reconciling information-seeking behavior with search user interfaces for the web.* Journal of the American Society of Information Science and Technology.

[4] *Hearst (2006). Clustering versus faceted categories for information exploration.* ACM, *49(4).*

[5] *Hearst, M. A. & Stoica, E. (2009). Nlp support for faceted navigation in scholarly collection.* Text and Citation Analysis for Scholarly Digital Libraries, *(pp. 62–70).*

[6] *Ka-Ping Yee, Kirsten Swearingen, K. L. & Hearst, M. (2003). Faceted metadata for image search and browsing.* In Procs. of CHI, *03.*

[K.W. Leung & Chung] *K.W. Leung, S. C. & Chung, F. Applying cross-level association rule mining to cold-start recommendations.* IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops.

[8] *Marchionini, G. (2006). Exploratory search: From finding to understanding.* ACM, *49(4), 41–46.*

[9] *P, P. (2008). Informing through user-centered exploratory search and human-computer interaction strategies.* Issues in Informing Science and Information Technology (IISIT), *5, 705–727.*

[10] *Senjuti Basu Roy, Haidong Wang, G. D. (2008). Minimum-effort driven dynamic faceted search in structured databases.*

[11] *Sieg A, Mobasher B, B. R. (2007). Representing context in representing context in web search with ontological user profiles.* Model Using Context*, 4635, 439–452.*

[12] *Spink, A., J. B. W. D. & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes.* IEEE Computer*, 35(3), 107–109.*

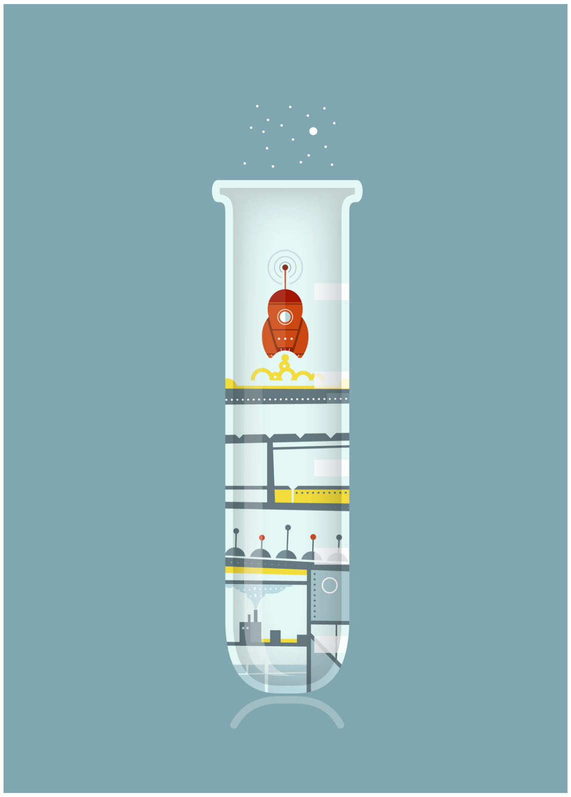
[13] *W. Gao, S. W. & Cerrone, N. (2002). A dynamic recommendation system based on log mining.* In International journal of foundations of computer science*, 13(4), 521–530.*

[14] *X. Shen, B. T. & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback.* In Proceedings of SIGIR*, 05, 43–50.*

[15] *Zaier, Z. (2010). Modèle multi-agent pour le filtrage collaboratif de l'information.*

[16] *Zhai, C. & Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. (pp. 403–410).*

[17] *Zhang, Y. (2004). Using bayesian priors to combine classifiers for adaptive filtering.* In SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 345–352).*



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.