An implicit feedback machenism of solving search behavior modeling

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.

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[1,13]. 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 key-

word 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[7]. 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[8]. 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[12] as the approach of Z. Zaier[14] 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[9]. 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[15,16]. 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

In this chapter, we first discuss the relationships between user's lower level and higher level behavior in web search background. This is a prelude for the rest of this chapter. We are intended to deal with the process of modeling user's higher abstraction level behavior and its implementation.

Our first task was to understand the space of user goals. In particular, we needed to come

up with a framework that could identify and organize a manageable set of canonical goal categories. These goal categories, in turn, must encompass the majority of actual goals users have in mind when searching.

2.1 Relationships of Lower Level and Higher Level Behavior

In the web era, search engines are used for more than just research. Even the most cursory look at the query logs of many major search engine makes it clear that the goal underlying web searches are many and varied. And while the vast body of work described above has helped us to understand *what* users are searching for and *how* their information-seeking process works, there have been few attempts to look at *why* users are searching.

One of the fear exceptions is Broder's "Taxonomy of Web Search"[3]. Motivated by the idea that the traditional notion of an "information need" might not adequately describe web searching, Broder came up with a trichotomy of web search "types": navigational, informational, and transactional. *Navigational* searches are those which are intended to find a specific web site that the user has in mind; *informational* searches are intended to find information about a topic; *transactional* searches are intended to "perform some web-mediated activity".

User's interests change over time, studies of user search behavior have a long history in Information and Library Science[2 11]. 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.

In order to definitively know the underlying goal of every user query, we would need to be able to ask the user about her interest. Clearly, this is not feasible in most cases[7]. But can the goal be determined simply by looking at the query itself, or is more information required? 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[10].

2.2 Traditional Faceted Search

In the traditional faceted search, each category is represented by a facet, user can choose different facets to drill down the search results. A screen shot of the classical faceted search is shown in Figure 2.1. Unlike document-based relevance feedback mechanism which asks users to give feedback on the relevance of documents. Faceted search allows users to give feedback on document metadata fields.

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.

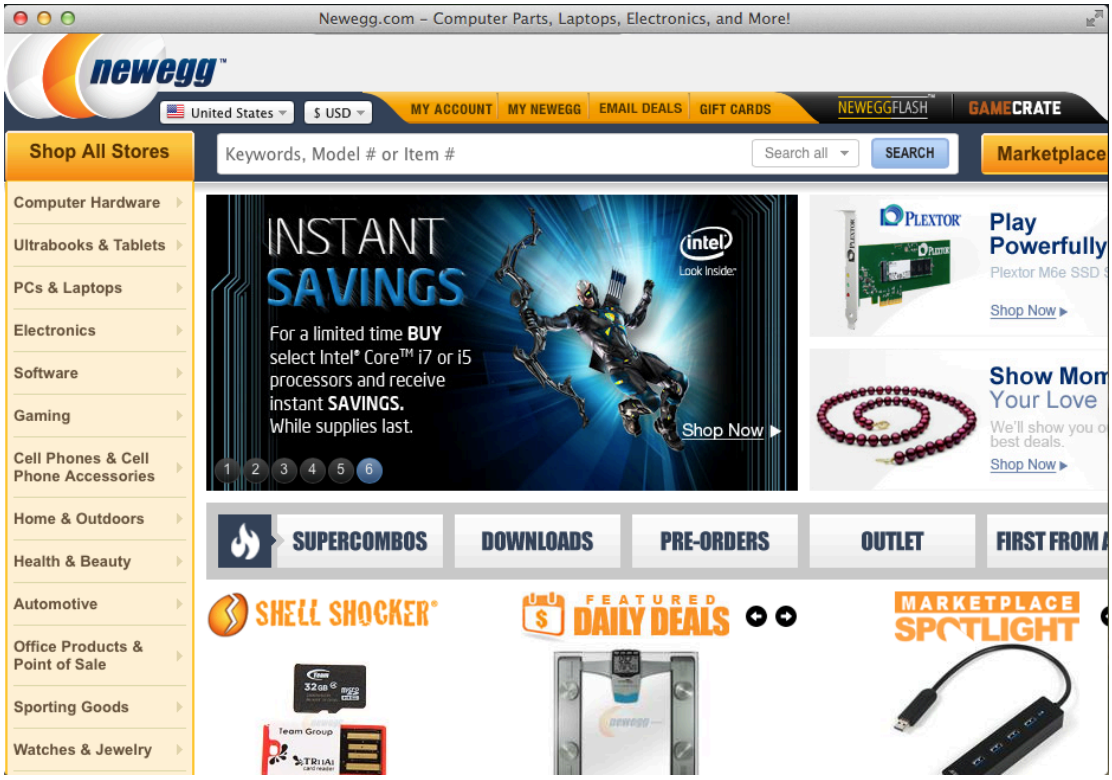

Figure 2.1: A screen shot of traditional faceted search.

2.3 Hits Queue Model for Facets

This common scenario inspires us to map user's clicks into a *Hits Queue*. In this paper, each metadata field is called a facet, and a facet (*f*) with a specific value (*v*) is called a facet-value pair (*f : v*). Each facet-value pair represents a faceted constraint on returned documents, E.g., language:Chinese, format:ppt: subject:IR, genre:comedy.

To void overwhelming users with many facet-value pair candidates, the system needs to recommend a small number of facet-value pairs that are most probably interesting to a user. A good recommendation approach is crucial in the faceted feedback mechanism. Intuitively, the recommended facet-value pair should be good in - they have a high score of being relevant and thus chosen by the user. Based on this respect, we propose *Top Document Score (TDS)* method.

To select the most frequent facet-value pairs occurring in the top N ranked documents returned by a baseline retrieval algorithm using the initial query. We calculate the interest score instead of frequency of each face-value pair in the top N documents, which is called "Top N Document Score". The top N most highest score facet-value pairs are chosen as candidates to present to the user. The underlying data structure of *Hits Queue* 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.

2.4 Mapping clicks to Hits Queue

A chosen *f* is presented by a facet in the priority queue. A *priority* contains a set of facet-value pair *f: v*. To map facet-value 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:

- Initialize each facet as *f: 0*.
- User clicks on a certain facet she is interested in.
- Improve this facet's score by 1, it becomes to the first node in the queue.
- User clicks on another facet.
- Reorder the queue with different priority. ...

The motivation of using *Hits Queue* is twofold: 1) a facet-value pair that appears rarely in the whole corpus while frequently in top ranked documents has a high probability to be relevant; 2) the retrieval system gets more benefits by knowing a rare facet-value pair covering a small number of documents being relevant that a frequent one.

3

The Search Results Model

In this chapter, we present a retrieval model to incorporate user faceted feedback. Comparing traditional flat facet mechanism that just supports drill down on a certain category, we purpose a graph model to represent search results . This is because the flat facet mechanism is based on two assumptions: 1) users are very clear about what they are looking for, and thus are able to select facet-value pairs to restrict the returned result; 2) document facets are

accurate and complete so that no potentially relevant document is filtered out in retrieval due to meta data errors. These two assumptions may not hold in text document retrieval.

3.1 Data Graph Model

In a specific domain, some facets might be more informative that others. For example, for new articles, the information of time, locations, persons, and topics may be more important than publishers; for research papers, the subjects and keywords may be more informative that the file formats; for movies, the genres, casts and directors may be more informative than producers.

Based on above motivations, we propose a data graph model. In this model, we learn a weight for each type of facet, which is expected to reflect the quality of the facet. Here, the quality may include user acquaintance, meta data accuracy, facet importance, etc.

Let D be a set of search results. D can either be a base relation or a materialized 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 data set. 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.

4

OnlyFacets exploratory search

This chapter deals with the process of exploratory search and its implementation. We purpose a prototype named *OnlyFacet* search, as to fully evaluate the usability of our search prototype, we intend to remove key-word search bar. To begin with, an introductory example is presented and the functionality of the prototype graphical user interface (GUI) is explained.

Figure 4.1: The *OnlyFacet* search GUI showing related entities for "American Place"

4.1 The user interface of *OnlyFacet* search

The graphical user interface (Fig)is designed to comprise three main areas: the direct search results in the center of column, the *faceted filter* on the right, and the exploratory search navigation on the left. The search results include a timeline, which shows the automatically generated temporal segmentation of the results including highlighted segments indicating search hits. The facet filter allows to narrow down the search results based on graph search.

Faceted search aims to broaden the scope of search by suggesting related terms, concepts and resources. Our approach uses priority queue and data graph to support the search process by exposing additional information about indexed resources.

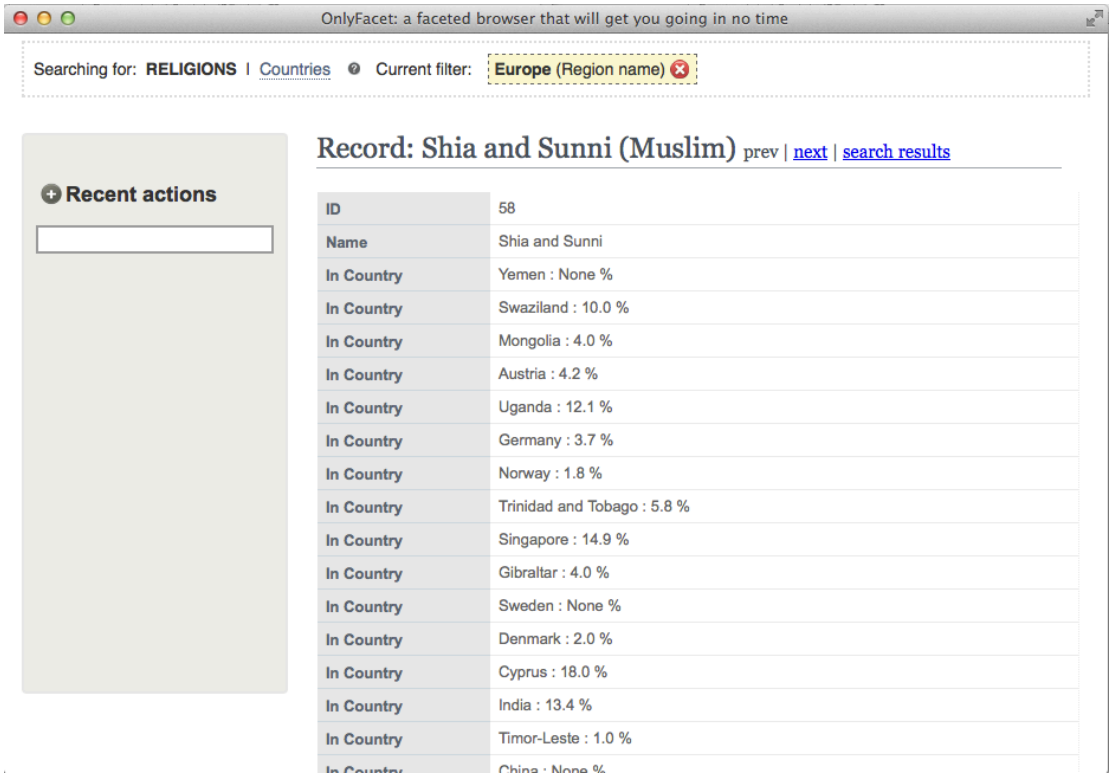

Figure 4.2: The *OnlyFacet* showing related entities for 'Shia and Sunni'

Figure 1 depicts the result of a query after the user clicking on the facets - restaurant and place. The exploratory search GUI suggests a list of related entities. When the user clicks multiple facets, the labels of the mapped entities grouped by their connecting properties.

By clicking on, e.g. *'Shia and Sunni'* in the faceted search GUI, a new search is issued and the GUI switches to the newly selected entity showing its related entities and properties (cf. Fig. 2). This supplementary information includes, as e. g., related places (birth place, work place, etc), predecessor and successor in the presidential office, or Barack Obama□s residence.

To retain previous actions, a history list (4) provides links to previous searches. Option-

ally, the user may activate an additional preview of the search results evoked by a related entity when clicking on it (5). Moving the mouse pointer over these previews causes a popup to show brief information about the video resource (6).

5

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 or-

dinary 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.

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).

[3] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, (pp.36).

[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.

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

[8] 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.

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

[10] 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.

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

[12] 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.

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

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

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

[16] 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).