Science of Computer Programming 94 (2014) 18–31

Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico

Approximately optimal facet value selection

Sonya Liberman [a,1], Ronny Lempel [b,*,2]

[a] CONTEXTin, Herzliya Pituach 46725, Israel
[b] Yahoo! Labs, Matam, Haifa 31905, Israel

HIGHLIGHTS

- We define a measure of user effort when searching in a faceted search system.
- We define two models for user drill-down behavior in faceted search systems.
- We show efficient approximation algorithms for minimizing search effort in our models.
- We show our algorithms reduce (simulated) user effort more than existing baselines.
- Our algorithms outperform baselines even when modeling assumptions do not hold.

ARTICLE INFO

Article history:
Received 1 November 2012
Received in revised form 19 June 2013
Accepted 30 July 2013
Available online 21 August 2013

Keywords:
Search engines
Multifaceted search
Facet value selection

ABSTRACT

Multifaceted search is a popular interaction paradigm for discovery and mining applications that allows users to digest, analyze and navigate through multidimensional data. A crucial aspect of faceted search applications is selecting the list of facet values to display to the user following each query. We call this the *facet value selection problem*.

When refining a query by drilling down into a facet value, documents that are associated with that facet value are promoted in the rankings. We formulate facet value selection as an optimization problem aiming to maximize the rank promotion of certain documents. As the optimization problem is NP-Hard, we propose an approximation algorithm for selecting an approximately optimal set of facet values per query.

We conducted experiments over hundreds of queries and search results of a large commercial search engine, comparing two flavors of our algorithm to facet value selection algorithms appearing in the literature. The results show that our algorithm significantly outperforms those baseline schemes.

1. Introduction

Multifaceted search, also known as guided navigation, is a popular and intuitive interaction paradigm for discovery and mining applications that allows users to digest, analyze and navigate through multidimensional data. Many Digital Libraries and e-commerce Web sites implement faceted search applications. Lately, facets have begun appearing also in result pages of general Web search engines. As Web queries are often short and ambiguous, facets can assist searchers in disambiguating their precise information need, or *intent*.

The interaction with a faceted search interface involves interleaved search and browse operations over semi-structured documents. In addition to containing text, the documents are associated with *facet values* – nodes (i.e. values) of a

* Corresponding author.
E-mail addresses: sonya_lib@yahoo.com (S. Liberman), rlempel@yahoo-inc.com (R. Lempel).

[1] Work done while interning at Yahoo! Labs, Haifa.

[2] This is an extended version of the SACWT'2012 paper by the same authors [1].

0167-6423/$ – see front matter
http://dx.doi.org/10.1016/j.scico.2013.07.019

multidimensional facet hierarchy/taxonomy. For brevity of notation, we hereby abbreviate "facet value" by *FV*. At any step in the search session the user may either (1) modify the search query, (2) browse (drill-down) into one of several displayed FVs that further *narrow* the context of the current query, or (3) remove some FVs from the context (roll-up), hence generalizing the context. Note that when narrowing a query by drilling down into a FV, search results are filtered to contain only those documents associated with the FV. Thus, results belonging to the clicked FV are *promoted* in the rankings, as documents that previously ranked ahead of them and are not associated with the FV, are filtered out.[3] Note that the result of a drill-down is very different than the result of modifying the query by appending the label of the FV as an additional query term, as (1) documents containing the term may not necessarily be associated with the FV taxonomy node; (2) documents associated with the taxonomy node may not necessarily contain the term; and (3) query terms affect results' ordering by complex scoring logic, whereas in most implementations drill-down operations act as rank-agnostic filters.

A crucial aspect of faceted search applications is selecting the list of FVs to display following each query. The role of the chosen FVs is to guide users toward satisfying their intents. Thus, the key is to display FVs that are well-aligned with users' (latent) intents, and that partition the information space in a manner that facilitates intent fulfillment via navigation. We call this the *facet value selection problem*. One can view facet value selection as the link generation process of the adaptive hypermedia navigation support [2] that drives faceted search applications.

This paper presents a novel algorithm for facet value selection. The algorithm selects FVs that optimize the rank promotion of documents it believes to satisfy users' intents. We model this belief by postulating a distribution over search results, assigning a "goodness probability" to each document. We then consider the promotion in rank of each document, attained by drilling down into each possible FV, and select a set of FVs that approximately maximizes the expected rank promotion of the documents. The mechanics of our approach guarantee a balance between selective FVs, which are associated with few documents but promote each such document significantly, and FVs that cover many documents and hence cannot promote them by significant amounts. Furthermore, our approach naturally favors the selection of non-correlated FVs, that promote the ranks of nearly disjoint sets of documents. Experiments conducted with two flavors of our algorithm show it is superior to facet value selection algorithms that appear in the literature.

The rest of this paper is organized as follows. We survey related work in Section 2. In Section 3 we model user interaction with faceted search applications, and formally define the optimization problem of facet value selection. Section 4 proves that the optimization problem is *submodular*, and proposes a greedy algorithm that selects a set of approximately optimal facet values. Sections 5 and 6 report on our experiments with the proposed new algorithms, and show that they outperform baseline algorithms that are referenced in the literature. We conclude in Section 7.

2. Related work

Faceted search applications require faceted data, namely the existence of facet hierarchies and the mapping of documents onto those hierarchies. In two related papers, Dakka et al. [3,4] describe algorithms for extraction of facet hierarchies from a corpus based on lexical subsumption, and assignment of the documents to those facets. Stoica and Hearst [5] use synsets and hypernym relations to accomplish a similar task. Feinstein and Smadja [6] describe the RawSugar social tagging system, which supports faceted search over tag hierarchies. Kohlschütter et al. [7] use personalized PageRank values for multiple ODP[4] categories to (1) infer dominant facets in Web search results, and (2) support drill-down operations on the result set. Anick and Tipirneni [8] map each document to a list of its terms of high lexical dispersion, and at query time display terms appearing in several top ranking documents.

Another crucial aspect in deploying faceted search is the user interface, whose purpose is to clearly present the multi-dimensional information space to users, guiding and enabling them to make informed decisions on their interleaved search and navigation steps within the space. Hearst [9] provides observations based on many years of experiments on interface design, most recently as part of UC Berkeley's Flamenco Search Interface project.[5] Kules [10] studies how organizing large result sets into categorized overviews helps users in exploring and understanding such result sets. Shen et al. [11] study visualization techniques in Digital Libraries that support integrated browsing and searching. Zhang and Marchionini [12] demonstrate that certain faceted UI choices can significantly shorten the time users require to complete search tasks.

Many faceted applications present aggregated statistics (e.g. document counts) for each facet value on its own. Beyond that, Schneiderman et al. [13] plot two-dimensional tables with *hieraxes* – axes of hierarchical categories. Meredith and Pieper's inverted index based BETA system [14] also displays two-dimensional tables for correlating pairs of facets values.

Between the data model and the user interfaces lies the index. The Apache Solr open source project[6] supports searching over a flat list of facets. Ben-Yitzhak et al. [15] describe an indexing scheme of hierarchical faceted data, along with the corresponding runtime algorithms for supporting various arithmetic and logical aggregations over the data.

Once the data model, the index structure and the UI paradigm have been decided and implemented, one reaches the main question addressed by this paper – which facet values are the most useful for every given query. In many commercial

[3] In this paper we further assume that the relative order of documents that are associated with the clicked FV does not change following the drill-down operation.

[4] Open Directory Project, http://www.dmoz.org/.

[5] See the Flamenco site http://flamenco.berkeley.edu/index.html for references to many additional publications.

[6] http://lucene.apache.org/solr/.