Structured Retrieval and User Profiles

Marijn Koolen[1] and Toine Bogers[2]

[1]Institute for Logic, Language and Compupation,
University of Amsterdam, The Netherlands ,
marijn.koolen@uva.nl
[2]Department of Communication and Psychology,
Aalborg University Copenhagen, Denmark,
toine@hum.aau.dk

Definition

The combination of structured information retrieval with user profile information represents the scenario where systems search with an explicit statement of the information need—a search query—as well as a profile of a user, which can contain information about previous interactions, search history, user demographics, or other relevant information about the user's preferences. The relation between the profile and the information need is implicit and may contain many irrelevant signals. The task of the system then is to model both the current information need and the background user preferences to derive notions of topical relevance as well as user relevance and to find the right balance between these notions to determine the optimal ranking of search results.

Historical Background

Information retrieval research has traditionally focused on locating documents that are relevant to a user's search query—an explicit statement of that user's underlying information need. However, it is well-known that a query is only a limited and compromised representation of the underlying need [**?**]. Which documents are relevant to the user is related to the user's context, which includes, among other things her background knowledge, interests and what the requested information will be used for. A user's previous interactions with the system can be captured in a user profile to build a model of her preferences and interests. Such a profile can consist of, among others, a list of viewed or selected documents, bookmarks, ratings, reviews, tags, and other signals of interaction.

Early work on combining user profiles with search queries was based on literature search, where a user's profile consisted of papers previously collected by

that user [? ?]. This model was also used for the TREC[1] Filtering Track [?], where the system has to filter a incoming stream of documents and track the topic(s) of interest to a user based on their previously selected documents. The TREC Contextual Suggestion Track [?], starting in 2012, addresses the related problem of suggesting activities for a user in a specific location, where there is no query but only a profile of user preferences. Just as the information retrieval community has investigated how to integrate user profile information into retrieval algorithms, the recommender systems community has long attempted to integrate content-based information and user profile information [? ? ? ?].

Due to the rise of social media, an increasing amount of information about the user can be exploited by search engines and recommender systems. Users have rich profiles on social network sites such as Facebook, LinkedIn and Google+ that reveal their interests, activities and their social circles. Other platforms allow users to rate, reviews and tag books, movies or songs, and recommend new items based on these interactions. These are domains with highly structured information in the form of metadata and user-generated content. Searching in such collaboratively created catalogs provides an interesting scenario to study the value of user profiles for retrieval.

One such platform is the social cataloging site LibraryThing (LT)[2], where users can build up a personal catalog of books and rate, review and tag these books or discuss them with other members. LT has a recommendation engine to support users in discovering new books, but users can also use directed search, which is the scenario that is investigated in the Social Book Search (SBS) tracks[3] at INEX and CLEF [?]. The SBS track investigates book search in a large collection of book descriptions containing professional metadata and user-generated content, with an elaborate statement of user information needs and an extensive profile of those users, all of which have rich structure that can be used to support retrieval [? ?].

Scientific Fundamentals

User profiles are typically used in recommender systems, where users rate items and the system derives implicit interests from these ratings to suggest other items to the user. These signals of interest may also help to contextualize the information need for directed search. However, few retrieval systems make use of such profiles to improve the ranking of retrieval results. Exploiting user profiles for directed search poses a substantial challenge. The profile contains signals of implicit interest and background knowledge, but how should these signals be combined with the explicit statement of the specific information need expressed by the search query? How is a user profile related to the search topic and which of a user's interests are relevant to that topic?

Formally, a user u has an information need represented by an query q and

[1] TExt Retrieval Conference, see also http://trec.nist.gov
[2] http://www.librarything.com
[3] http://social-book-search.humanities.uva.nl/

Fig. 1: Book request on the LibraryThing forum

related interests and background knowledge represented by a user profile p_u and the challenge is to rank a collection D of documents d by relevance $R(d|q, p_u)$ such that the most relevant documents are ranked highest. Documents and profiles can have a rich metadata structure, allowing for complex retrieval and ranking models that carefully balance the evidence from q, p and d [**?**].

The SBS evaluation campaign studies this problem in the context of the LibraryThing discussion forums, which are used to discuss a broad range of book-related topics. Many members turn to these forums asking for book suggestions with other members replying and providing suggestions. The LT discussion forums provide a unique opportunity to unobtrusively investigate complex, realistic search scenarios with highly structured data in the form book metadata and user profiles. At the same time, the data from the forums is used to construct test collections for developing and evaluating search systems that combine structured retrieval with user profile information. Examples of book metadata that could be exploited for structured retrieval are book titles, author names, subject descriptors, product descriptions and user tags and reviews. User profiles also contain personal tags, ratings, and reviews. The result of so many structural elements is a variety of perspectives on a book from different types of agents (authors, publishers, library catalogers and readers), and for a broad range of information needs—combining aspects of topic, genre, style, mood, time and interest. The challenge is then to identify and weight the strongest signals of relevance in this abundance of evidence.

The request in Figure **??** is highly complex, providing requirements about the content, examples of books and authors that the poster of the request is already familiar with, and contextual cues on usage. The user name links to

Fig. 2: Book request on the LibraryThing forum

their profile, which provides contextual information on the user (Figure **??**, including the books they cataloged and how they tagged and rated them. The example books introduce a form of querying-by-example that could also be seen as a recommendation task. Other forum members reply in the thread with book suggestions relevant to the request. These suggestions represent human relevance judgments and are used, in combination with the user profile of the topic creator, to evaluate retrieval systems. The user profile of the requester contains information on when they added each book to their catalog and the forum thread has timestamps on when the request was posted and when suggestions were made. Book suggestions in the forum thread that are cataloged by the requester after posting the request are considered more relevant than suggestions that the requester ignores or already cataloged before the request. This provides a complex picture of relevance that goes beyond topical relevance, touching on situational and user aspects of relevance.

Key applications

There are several domains where search engines could make use of rich document structure and contextual information from user profiles, such as travel, online shopping, scientific literature.

Travel Users who are organizing a trip and are searching options for travel, accommodation and activities. This is a complex search task where specific requirements and personal preferences need to be balanced and the available data is often highly structured (names, locations, dates, prices, ratings, reviews). Information about past travel plans could help direct the search process towards the most optimal arrangements.

Online shopping Shopping on e-commerce websites such as Amazon.com often involves complex information needs that require multiple searches to locate the most appropriate products. User profile information could help to personalize the search results and provide more relevant product suggestions.

Literature search There is rich structured information that could be exploited when a researcher searches for relevant literature to reference in their paper. The researcher's interest and background knowledge can be represented by their own publications and the publications they reference. Each individual publication has metadata such journal or proceedings title, year, keywords and references [**?**] .

Experimental Results

A well-established result in book search is that user-generated content is very effective for retrieving relevant suggestions for a range of tasks [**?** **?**]. The most effective approaches use a mixture of models for different types of metadata, essentially exploiting the principle of *polyrepresentation* (representing the same information by data from multiple sources).

Forum suggestions are different from traditional topical relevance judgments [**?**]. Forum members take into account writing style, engagement, humor or how comprehensively a topic is discussed. Members have read most of the books they suggest in response to a request and mention books mostly positively [**?**].

The best performing systems use all available evidence for the information need (represented by different fields such as thread title, discussion group and text of the first post), as they all provide different perspectives on the information need, again exploiting the principle of polyrepresentation.

The user profiles have been effectively used for combining retrieval results with collaborative filtering recommendations, where books are recommended if they are prominent in the catalogs of users similar to the requester [**?**]. More recently, **?**] rank books from similar users and find that precision is low, but that recall goes up linearly after rank 500 whereas runs based on textual queries gain almost no recall that far down the results list.

Data Sets

Data sets in IR are typically test collections constructed in evaluation campaigns like TREC, CLEF, NTCIR and INEX. Test collections with structured document collections and user profiles are available from the CLEF Social Book Search Lab[4]. The SBS test collection consists of a set of documents, user profiles, search topics and relevance judgments. The document set contains 2.8 million book description with professional metadata from Amazon—such as title, author, publisher and subject descriptors—and user-generated content from Amazon (user reviews and ratings) and LT (user tags). The search topics are made up of LT forum topics that are focused on a request for book suggestions

[4] Available at http://social-book-search.humanities.uva.nl/.

and consist of the thread title, the initial post and discussion group name. The relevance judgments are based on the forum suggestions by other LT users. In addition, there is a set of 94,610 user profiles containing the books that each user cataloged, the cataloging date and user ratings and tags. Another relevant test collection containing structured document collections and user profiles is that of the TREC Contextual Suggestion tracks [5]

Cross References

STRUCTURED DOCUMENT RETRIEVAL

SEARCH TASKS

INFORMATION RETRIEVAL

[5] Available at https://sites.google.com/site/treccontext/.