Workshop on Aggregated Search

Vanessa Murdock
Yahoo! Research
Diagonal 177 Planta 8
08018 Barcelona Spain
vmurdock@yahoo-inc.com

Mounia Lalmas
Department of Computing Science
University of Glasgow
Sir Alwyn Williams Building
Lilybank Gardens
Glasgow G12 8QQ UK
mounia@acm.org

Abstract

The Workshop on Aggregated seeks to define the research problems for aggregated search results, and to build a community of researchers working in this emerging area. Aggregated search addresses the issues of how to select and rank resources, how to present them to a user in such a way that information can be found efficiently and effectively. The aim of the workshop was to encourage discussion and the exchange of ideas about the directions for aggregated search. The workshop presented prototype systems, user studies of existing systems, as well as papers about component technologies.

1 Introduction

The search paradigm by which a user types in three or four words for a query, and receives a ranked list of results, becomes less effective when the information the user seeks is not contained in a single document, or even in a single category of resource. Furthermore, such an interface which requires a user to scan a ranked list, and click the result and then search the document for the information, is cumbersome if the user is using a mobile phone, or hand-held device. As this is an emerging area of interest, in a fast-growing and dynamic environment, there are a number of directions search technology might take. Aggregated search is the task of searching and assembling information from a variety of sources, placing it in a single interface.

If we do not assume the user will interact with the system by typing in a set of keywords, and we do not require the presentation of a ranked list of documents, we open up questions about satisfying an information need that have never been asked before.

- *What is the core information the user is seeking?* Previous work on passage retrieval, XML retrieval, question answering, high-accuracy retrieval of documents, and summarization all attempt to identify the information need of the user, and find the exact information the user seeks. But there are other questions, such as identifying based on

the query and the search context that the best information is a photo, a map, a video clip, or an mp3 file, often in addition, or as an alternative to text results.

- *What information should be presented to the user?* If we do not limit the results to a ranked list of documents, then the information itself must be assembled in some way. If a person has indicated they are looking for restaurant recommendations, then perhaps they need a list of restaurants, and blog entries or reviews of the restaurants, along with a map, and perhaps a menu.
- *How should the information be presented to the user?* It is not clear that a ranked list is an effective interface for users of mobile devices, or for people seeking non-textual information. In this case, there is a question of the best interface that presents the information in the most usable form, and allows the user to interact in the most natural way possible. Depending on the device used, perhaps the best interface is a single page constructed from the constituent information, or perhaps it is a set of pages that the user can flip through like a book, or a pamphlet.
- *How should the user interact with the system?* At the present time, there is an assumption that the user will interact with the retrieval engine by entering textual queries. Some research on image recognition aims to allow the user to query with a photo. Work in natural language processing and interactive retrieval aims to allow the user to interact in a natural language interface. Further, research in multi-modal interactions allows the user to speak, or gesture with a pointer device.

Examples of aggregated search include Alpha Yahoo[1] and Universal Search by Google[2]. A single query yields results from a variety of vertical searches, including image, video, community question answering sites, news articles, and sponsored results. The challenge is to decide what vertical properties to present results from, and how to organize the results for the user. Furthermore, although Alpha Yahoo and Universal Search are restricted to textual queries, it need not be the case as the uses for search expand into a wider range of devices.

Of the research in aggregated search, much has centered on structured retrieval, such as XML retrieval. In XML retrieval the aim is to identify the most specific elements to return in answer to a query, whether or not the query is expressed with structural constraints. It has been shown that returning several elements together as one answer triggers stronger user satisfaction than returning a single element. In the "relevance in context" retrieval task[3], the aim is to return documents constructed from only the most specific and relevant elements from the original document. In this case, the result is aggregated, but elements were selected from the one document. In the more general setting, elements are selected from different documents to form an aggregated result (a query language like XQuery Full-Text could be used to specify how), and the documents need not be textual in nature.

The workshop aimed to build a community of researchers to begin to define the problem and tackle the questions inherent in such a search paradigm. The workshop attracted around thirty participants, with five paper presentations, one demo, two invited talks, ending with a joint panel with the Workshop on Focused Retrieval.

[1] http://au.alpha.yahoo.com/

[2] http://searchengineland.com/070516-143312.php

[3] Part of INEX, the evaluation initiative for XML retrieval, http://inex.is.informatik.uni-duisburg.de/

2 Presentations

The first invited speaker, Soyeon Park, presented a study of user behavior in a major commercial search engine, Naver.com, which was the first commercial aggregated search system. The second invited speaker, Hugo Zaragoza, addressed practical issues in constructing synthetic documents, with an aggregated search prototype *Yahoo! Correlator* built on top of the English Wikipedia corpus. Paper presentations covered a wide range of topics related to extracting and organizing information, and results presentation.

Soyeon Park, from the Duksung Women's University and Joon Ho Lee, from NHN Corp., conducted a study of user behavior on Naver.com. Naver.com launched the first commercial aggregated search engine in 2000. Aggregated search pages are typically more information rich than traditional ranked lists. This is because they present many more results from multiple genres, in a more compact format. Such a user study is vital to understand both how to select the information to be presented, and to organize it in a way that users can find information efficiently. Some of the results of the user study indicate that users typically enter short queries, seldom using advanced search features, and viewed few result pages. Furthermore users are passive, seldom making changes to the search environment set by the system. In Naver, the aggregated or "unified" search page is the default, and nearly all users used the unified search. When queries are classified as more suitable for a particular collection, most users still used the unified search. In addition to the user behavior, Park also presented trends in terms of topics users search for, and discussed how the collections are ranked, as well as how the documents within each collection are ranked.

Hugo Zaragoza from Yahoo! Research presented a demonstration of an aggregated search prototype *Yahoo! Correlator* built on the English Wikipedia corpus. Correlator constructs a document on the fly in response to a user query from parts of other documents in the corpus. The demonstration focused on different types of documents that might be constructed for a variety of queries. The technology behind the synthetic documents uses an efficient indexing and querying system that allows for linguistic entities to be extracted, and ranked in response to a query. The system does not depend on entities being tagged in the text or listed in a thesaurus, and the system is designed to be scalable. In their system, sentences are indexed by a passage-retrieval engine, and an entity-containment graph is constructed to represent the entities in the text. Sentences ranked both by their relevance to the query, and according to the entity-containment graph, so that sentences that best explain the importance of the entity with respect to the query are ranked higher. A user-interface displays the selected sentences using an entity-dependent layout. For example, geopolitical entities can be drawn on a map, whereas temporal entities can be shown on a time-line.

The papers covered a wide variety of topics, and due to space considerations we cannot summarize them here. The papers presented were:

- *Aggregating Search Results for Social Science by Extracting and Organizing Research Concepts and Relations* by Shiyan Ou and Christopher S. G. Khoo
- *Bibliometric Maps for Aggregated Visual Browsing in Digital Libraries* by Andreas Strotmann and Dangzhi Zhao
- *Using Digest Pages to Increase User Result Space: Preliminary Designs* by Shanu Sushmita, Mounia Lalmas and Anastasio Tombros
- *Aggregated Search: An Indian Experiment* by Ravindra Dastikop, Parashwanath R. Tadas, and Govindappa A. Radder

- *Leveraging Query Associations in Federated Search* by Aditya Pal and Jaya Kawale
- *From Aggravated to Aggregated Search: Improving Utility through Coherent Organisations of an Answer Space* by Stephen Wan, Cécile Paris and Alexander Krumpholz

3 Panel Discussion

The workshop ended with a joint panel with the Workshop on Focused Retrieval "Beyond Document Retrieval: Zooming In, Zooming Out". The members of the panel were Jaap Kamps (Chair), W. Bruce Croft, Djoerd Hiemstra, Peter Ingwersen, Ray Larson, and Cécile Paris. Each panel member was asked to address the question of whether whole documents are the most appropriate unit of retrieval in all contexts. Is there value in zooming in to a document to access a piece of information directly, or zooming out to provide an overview of the relevance of multiple documents.

Peter Ingwersen equated zooming out with an emphasis on recall, and zooming in with an emphasis on precision, and advocated for a structured interface rather than an interface resembling the typical ranked list, and that doing so allows us to communicate more complex structure and answers. He brought up the point that the temporal dimension in queries is an important issue, as is the user interface and how the user interacts with the system.

Bruce Croft pointed out that to a certain extent the discussion around aggregated search and focused retrieval apply a new terminology to familiar technologies. For example, focused retrieval covers technologies such as sentence and passage retrieval, and question answering. He also pointed out that in discussing focused retrieval and aggregated search, we should define what are the specific research topics that are not yet solved.

Open problems identified by the panelists and audience members include how to define the appropriate granularity of an answer, how to evaluate the results, and whether to combine results in a non-vertical way, or to present a single ranked list of aggregated results.

Cécile Paris asserted that the research issue in aggregated search is how to combine results in a coherent way, depending on the task at hand. For aggregated search, she pointed out that relevance is not sufficient, that the results much help the user solve the problem, and the entities must be related in a meaningful way. That is, the interface is dependent on the task and the context of the search. She also pointed out that this is an area in which discourse structure and computational linguistics would be helpful.

Ray Larson advocated for looking at the document created as a whole rather than as a collection of snippets, and suggested a portal interface might be created for a given query.

Djoerd Hiemstra suggested that we would like to select documents that have the highest probability to be popular in the future, and pointed out that services like Alpha Yahoo are not really aggregated search as they do not combine actual results, rather they present results from multiple verticals in a single interface without making decisions about the granularity of the search.

4 Acknowledgments

We would like to thank the organizers of the Workshop on Focused Retrieval for helping to organize the joint panel discussion. We would like to thank the invited speakers, and Ricardo Baeza-Yates and the Yahoo! Research Lab in Barcelona for sponsoring them.