Inf Retrieval (2010) 13:407–411
DOI 10.1007/s10791-010-9137-5

Current research in focused retrieval and result aggregation

Andrew Trotman · Shlomo Geva · Jaap Kamps · Mounia Lalmas · Vanessa Murdock

Received: 11 May 2010 / Accepted: 24 May 2010 / Published online: 11 June 2010

Abstract Both focused retrieval and result aggregation provide the user with answers to their information needs, rather than just pointers to whole documents. Focused retrieval identifies not only relevant documents but also which parts of those documents are relevant, thus reducing the time it takes the user to navigate in a document. Result aggregation is used when the best way to fulfil the user's need is to draw from many different information sources (different collections, documents, or parts of the same document), and to aggregate these into a single result, thus reducing the time it takes the user to fulfil their information need. This special issue includes seven papers showing the breadth and depth of the current state of research in these two branches of information retrieval. They include descriptions of live result aggregation systems, and experimental focussed retrieval systems including sentence retrieval, question answering, and entity ranking; as well as metrics for passage retrieval. To introduce this special issue, we provide an overview of the work presented in these papers.

Keywords Focused retrieval · Result aggregation · Metrics · Users

A. Trotman (✉)
University of Otago, Otago, New Zealand
e-mail: andrew@cs.otago.ac.nz

S. Geva
Queensland University of Technology, Brisbane Qld, Australia

J. Kamps
University of Amsterdam, Amsterdam, The Netherlands

M. Lalmas
University of Glasgow, Glasgow, UK

V. Murdock
Yahoo! Research Barcelona, Barcelona, Spain

1 Introduction

Standard document retrieval finds atomic documents. It leaves to the user the task of locating relevant information within a document. It also leaves to the user the task of aggregating different results (each corresponding to a piece of the sought information) into the final answer. Focused retrieval addresses the first task by providing the user direct access to relevant information; result aggregation addresses the second task by a creating a single "result", an answer constructed from the relevant components.

Focused retrieval aims to identify not only documents relevant to a user information need, but also where within those documents the relevant information is located. It aims to satisfy the user's information need and not to just identify documents that satisfy the information need. There are four main forms of focused retrieval: element retrieval, passage retrieval, question answering, and entity retrieval. Element retrieval (also known as XML-IR) can be applied when the documents in the collection contain some kind of markup (such as XML). The retrieval engine will typically exploit the structure to identify the most relevant paragraphs, sections or documents to return as answers to a query. With passage retrieval the retrieval engine will typically choose the appropriate size of results to return and the location, based mostly on the content of the document (and sometimes its structure). Whereas element retrieval and passage retrieval are used for information seeking questions, question answering (QA) and entity retrieval aim to answer more fact seeking questions, and makes use of natural language processing techniques.

Result aggregation is a form of automatic document construction. Given a set of documents and document components that satisfy a user's information need (perhaps identified using focused retrieval), an aggregator will combine these into a single result. Techniques include multiple-component summarization, meta-search like result presentation, and mixed-media presentation (when searching over heterogeneous collections of, for example, text, images, video, and music).

The focused retrieval paradigm holds that the best result to present to a user is contiguous passages from a single document. That might be single word answer to a question, a list of entities, or a series of passages. In result aggregation this restriction is lifted. The user has an information need, and the document collection is an information repository. The problem is aggregating the information extracted from any number of passages of any number of documents into a single coherent result. This draws on many aspects of computer science including natural language processing, information retrieval, and human computer interaction.

Focused retrieval and result aggregation are two fascinating research avenues in information retrieval. As the papers in this special issue will attest, there are ingenious real-world applications of search and retrieval technology pertinent to and building up from many interesting research areas. A brief summary of the papers in this special issue follows.

2 Result aggregation

Kaptein and Marx (2010) examine focused retrieval and result aggregation in a collection of Dutch parliamentary transcripts covering more than 40 years. They first convert the transcripts into XML from which they can perform focused retrieval. In addition, they are also able to aggregate results in many different ways. As a parliamentary session is a sequence of utterances, they are able to take a document (a parliamentary debate) and graphically render these utterances on a time line. The user is able to see who spoke when and for how long. When a user clicks on the timeline they are taken directly to that utterance; the timeline is hypertext linked directly into parts of the documents. From just a timeline it is hard to determine who is arguing with whom. For this they introduce interruption graphs. Speakers are nodes and the edges between nodes represent an interruption of one person by the other. The thickness of the edge is a function of the number of times the interrupter interrupted. To summarise utterances they draw word clouds. Such clouds can be drawn for a single utterance, a speech, or a debate.

Paris et al. (2010) discuss the scenario in which the query cannot be answered by a single fact, web link, or data source. An example is obtaining information to plan for a trip; where the answer could be a customised travel guide. With current search technology the user must issue individual queries to many different data sources, obtain many separate pieces of information (general information, flights, hotels, and so on), and then manually aggregate the results. The automatic organisation of information into a coherent whole can save the user considerable time. A good aggregation will help the user make sense of how each contributing piece of information relates to each other, as well as to the whole. Paris et al. draw on research in natural language generation and information retrieval. They suggest that combining both areas of research will be particularly fruitful when dealing with complex information needs. Three concrete examples of how an aggregated search system would produce an answer are given. These examples are drawn from applications developed by the authors in the surveillance domain, the tourism domain, and in an enterprise setting. They have, for instance, a system that produces on-the-fly brochures about their enterprise.

3 Focused retrieval

3.1 Performance measures

Focused or sub-document retrieval aims to guide searchers directly to the relevant information in documents. This, in turn, leads to systems that guide how the user browses and reads the document. Arvola et al. (2010) address the problem of evaluating this reading order. This is of particular importance in resource-constrained settings (such as PDAs and smartphones) or when result lists include condensed summaries of documents. In these scenarios the reading effort is a top concern since a searcher cannot skim-reading, especially if navigation is difficult. The authors develop evaluation measures that take the user's reading effort directly into account in a "fetch and browse" scenario. Their work contributes to the current discussion on utility-based measures versus effort-based measures, and on user models underlying measures. The explicit assumptions on user navigational behaviour underlying these measures greatly facilitate deciding whether the measures are applicable in a given search scenario, how they differ from other measures, and how to refine them in case a different user model is assumed.

3.2 Sentence retrieval

Sentence retrieval is important to many retrieval scenarios, e.g. summarization and question answering systems. As sentences are substantially different from documents, it has already been shown that models developed for document retrieval perform poorly on sentence-length texts. This is because sentences contain few terms, and the effects of the vocabulary mismatch problem are exacerbated. Losada (2010) addresses the vocabulary mismatch problem for sentence retrieval by comparing query expansion with pseudo-relevance feedback; and comparing sampling the expansion terms from the document context of the sentence (before retrieving sentences), with sampling from the set of top-ranked sentences (after the sentence retrieval step). Because sentence retrieval is a core technology upon which other applications are built, one application might require high precision, whereas another high recall. Losada's work examines the balance between precision and recall for each of his models. He advances the state of the art with an in-depth analysis of the conditions under which the local context of the sentence is useful, when to use pseudo-relevance feedback compared to query expansion, and from which context to sample expansion terms. The result is a concrete set of recommendations, grounded in empirical results, and a greater understanding of the nature of sentence retrieval.

3.3 Question answering

Moriceau and Tannier (2010) describe FIDJI, a Question Answering system for French. A search engine is used to identify whole documents and FIDJI then post-processes these documents, finds candidate answers to the question, and selects the answers that can be validated by syntactic analysis of the prospective answer text. Validation is performed by deriving dependency relations from the question and trying to match these to dependency relations derived from the candidate passages. FIDJI is an attractive system because it has a small footprint—it does not rely on syntactic analysis of the entire collection, nor does it rely on auxiliary ontologies or dictionaries. Instead it utilizes a search engine and then performs syntactic analysis on only a limited set of top documents. On the web scale, pre-processing of the entire collection is prohibitive but the lightweight nature of FIDJI makes web QA possible. But there is a down-side, run-time text analysis adds to query execution time; but the approach is independent from the underlying search engine.

3.4 Entity retrieval

Demartini et al. (2010) cast the entity retrieval problem as matching RDF triples (consisting of a subject, a predicate, and an object) by treating entities (subjects) as sets of attribute-value (predicate-object) pairs. This has the attractive property that various information sources can be combined seamlessly. In Wikipedia, pages of entities tend to have info-boxes with attribute-value pairs; but the page itself can also be regarded as a description (so a value of the "page" attribute). The flexibility of this approach facilitates the in-depth exploration of the relative value of different types of information for entity retrieval and list completion. The paper contains a comprehensive analysis of the effectiveness of different query representations (keyword query, entity type, example entity), document representations (document text, links, categories), and various NLP techniques (lexical compounds, synonyms and related terms, syntactic categories, named entities, semantic annotations).

They provide a wealth of detailed conclusions (both positive and negative) on the effectiveness of various techniques in isolation and in combination. For example, knowing the entity's type in terms of a relevant Wikipedia category is a highly effective cue for most topics, but should be used with care since a poor choice by the topic author can lead to bad results: there is considerable performance variation per query. The overall conclusion is that entity retrieval is a challenging problem on the Wikipedia, and even more challenging on the Web. It requires methods that can deal well with imperfect and uncertain cues.

Pehcevski et al. (2010) describe their entity ranking system also for use with the Wikipedia. To determine the final rsv for an entity (a Wikipedia document) they combine three scores, a document score, a Wikipedia category score, and a link score. They use BM25 for the document score. They look for semantically close categories to those seen in query examples. And with respect to links they show that using local context outperforms using the whole document's link structure. They go on to explore query difficulty using a considerable number of features taken from the topic and from the initial results list (such as the number of links in the top documents). They find a similar result to that seen in ad hoc query difficulty prediction – no evidence that such techniques can be used predictively to improve performance. They did, however, notice that in easy topics the category scores are important, whereas in difficult topics the link structure scores are important.

4 Conclusions

The aim of focused retrieval is to identify relevant information within a document. In doing so a focused system tries to fulfil the user's information need rather than just find documents. The size of the result is dependent on the query. The sentence is the ideal result size when aggregating many results into a single page. A single word or short phrase could answer a factual question. A list of entities would be the natural response to queries based on entities or entity types.

The aim of result aggregation is to move beyond the presentation of a single document as a result. In doing so an aggregating system tries to fulfil the user's information need rather than just present documents. Such system might either aggregate many documents into one (or one results page), or aggregate different parts of a single document so that the user can better understand the content of that document.

Used together, it becomes possible to move well beyond the whole document paradigm that has been at the centre of information retrieval research for many years. It is possible to extract relevant information from within documents, to aggregate it, and to fulfil the user's information need—rather than just give them a list of documents. Such systems are more than a hypothetical research agenda; live systems already rely on this technology. The seven papers in this special issue discuss such systems as well as research in both focused retrieval and result aggregation.

References

Arvola, P., Kekäläinen, J., & Junkkari, M. (2010). Expected reading effort in focused retrieval evaluation (this volume).

Demartini, G., Firan, C. S., Iofciu, T., Krestel, R., & Nejdl, W. (2010). Why finding entities in Wikipedia is difficult, sometimes (this volume).

Kaptein, R., & Marx, M. (2010). Focused retrieval and result aggregation with political data (this volume).

Losada, D. (2010). Statistical query expansion for sentence retrieval and its effects on weak and strong queries (this volume).

Moriceau, V., & Tannier, X. (2010). FIDJI: Using syntax for validating answers in multiple documents (this volume).

Paris, C., Wan, S., & Thomas, P. (2010). Focused and aggregated search: A perspective from natural language generation (this volume).

Pehcevski, J., Thom, J.A., Vercoustre, A.-M., & Naumovski, V. (2010) Entity ranking in Wikipedia utilising categories, links and topic difficulty prediction (this volume).