High-Precision Semantic Search by Generating and Testing Questions

W.J. Black, C.J. Rupp, C. Nobata, J. McNaught, J. Tsujii and S. Ananiadou
NaCTeM, School of Computer Science, University of Manchester

Introduction

This paper presents a novel way of allowing users to query large collections of scientific papers, looking for facts rather than documents. From ordinary search terms, queries are generated and tested, and only those known to match facts expressed within individual document sentences are proposed to the user, who then selects them from a pull-down list. A fact retrieval service aims to serve the research process efficiently by firstly searching for facts rather than keywords, and secondly by highlighting the facts in document snippets presented in the overview of search results.

Fact retrieval services providing authoritative answers to questions can be based on small databases or knowledge bases of manually curated information, or on document collections with manually curated annotations, but for huge document collections such as PubMed[4], the approach must be capable of automation.

We base our approach on MEDIE [2], where the user can express queries using a fact template with three slots, for the subject, verb and object of an elementary factual sentence, as shown in Figure 1. Output of query responses, illustrated below the input, presents the snippets from matching documents at the sentence level. If the user wants to see the wider context of the abstract, they can switch the view from *table* (shown) or *sentence* to *article*.

Indexing for Fact Retrieval

The MEDIE service has indexed the whole of PubMed abstracts, whereas our service is for the smaller full-text document collection managed by UKPMC. Both the MEDIE index and that for the FactFinder are generated by first parsing the full text of the source data using the Enju parser, which gives a full deep analysis of each document sentence. From this point on, the two approaches differ. MEDIE provides a query language that matches parse tree elements directly, whereas the FactFinder runs an offline process that transforms parse tree fragments into an index that can be directly managed by the Solr search engine framework. The two approaches have different advantages. MEDIE's query language means that power users can go beyond the template of Figure 1 and form complex queries that can match compound facts. On the other hand, the approach taken in the FactFinder has the merit of requiring a smaller footprint, using fewer computing resources at query time.

The fact index has one entry per fact (and not one per document, although the fact indices are qualified by their parent document id). The syntactic roles played by phrases within the sentence in expressing a fact are the fields in the index. These fields include subj, pred, obj, and represent the logical subject, predicate and object of the sentence, respectively. There are also fields naming the semantic class (e.g. protein, disease, drug, metabolite,...) of the domain terms within the facts.

The fact index may be relatively sparse, as we do not index every fact or assertion in the whole text, only those that express a biologically relevant fact. (e.g. sentences about method will typically be omitted as a result of our approach). A biologically relevant fact is taken as one whose pred is a verb from the BioLexicon, and at least one of whose logical arguments (subj, obj, etc.) is annotated as one of the semantic classes that our named entity indexer has annotated.

Query by question generation

For integration into an existing full-text document retrieval service, an alternative style of query form such as that in Figure 1 would sit uneasily alongside the normal expectation that a search interface consists of a text input box and a "search" button for input and a paginated ranked list of document summaries as output.

Helped by the observation that many search services offer suggested queries, typically comprising sets of query terms including those already input, we decided to see if we could generate semantic query suggestions that would match the facts in our fact index.

Figure 2 shows an illustration of an early prototype of the fact finder service, on which we have superimposed a full-text query results listing from the existing UKPMC service. All the questions proposed in the drop-down list are known to have "answers" in the fact index, the number being indicated in parentheses. The English query "*What associates with diabetes mellitus?*" corresponds to the formal query `+pred:associate +pmod_with:"diabetes mellitus"`, for example.

The process of query generation uses the facet query capability of the Solr search service. Facet querying [1] is intended to allow users to break down query result sets into subsets according to properties of the individual results. This is most familiar in search applications related to eCommerce, where users can break search results into groups by price range and product features, although Kleio [3] is a PubMed search service running facet queries based on the same named entity annotations as the FactFinder.

Facet queries return sets of values found for different fields in the records indexed by the user's original query term(s). We use these results to suggest alternative permutations of variable instantiations for each of a set of question templates. Each of the fully-instantiated questions is then run against the fact index, and those that have answers are posed to the user as query suggestions in the syntax of an English question.

When a user enters a query, in the main search box, it is handled as a regular full-text query, with a paged set of query results returned to the user. At the same time, the query string is also sent to the "fact query suggestion" service, which returns a set of possible queries against the index of facts. When these are received from the Web service, they are presented as a list that the user can select from. When a proposed fact-finder question is selected, the query results are extended with matching fact extracts of single sentences, to come within "fair dealing" limits.

Current status and future plans

The prototype shown in Figure 2 was a pilot implementation designed to test the technical feasibility of the approach. As a result of discussions with the project partners, various user interface changes have been requested which are planned for inclusion in a prototype to be tested with a focus group in July 2010. For example, the suggested questions will be ranked according to the size of their answer sets, and that the second input box shown in the figure will be removed. The focus group will be asked to comment on the value of highlighting the facts matching the chosen query in various different ways. Current technical work is focussed on refining the question generation process. The service is scheduled for deployment as a Beta service towards the end of 2010, and to go live in early 2011.

References

[1] M. Hearst. UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. In *Workshop on Human-Computer Interaction and Information Retrieval, HCIR 2008*, Redmond, WA, 2008.

[2] Y. Miyao, T. Ohta, K. Masuda, Y. Tsuruoka, K. Yoshida, T. Ninomiya, and J. Tsujii. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of COLING-ACL 2006*, pages 1017–1024, Sydney, Australia, 2006.

[3] C. Nobata, P. Cotter, N. Okazaki, B. Rea, Y. Sasaki, Y. Tsuruoka, J. Tsujii, and S. Ananiadou. Kleio: a knowledge-enriched information retrieval system for biology. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 787–788, Singapore, 2008. ACM.

[4] PubMed, 2007. http://www.pubmed.gov/.

Appendices

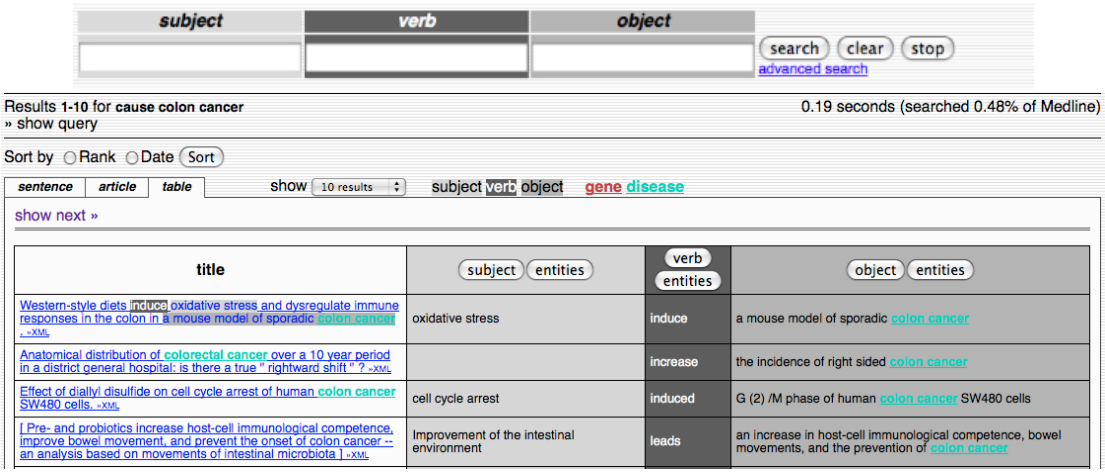

Figure 1: MEDIE's input form for fact queries and tabular display of matching sentences

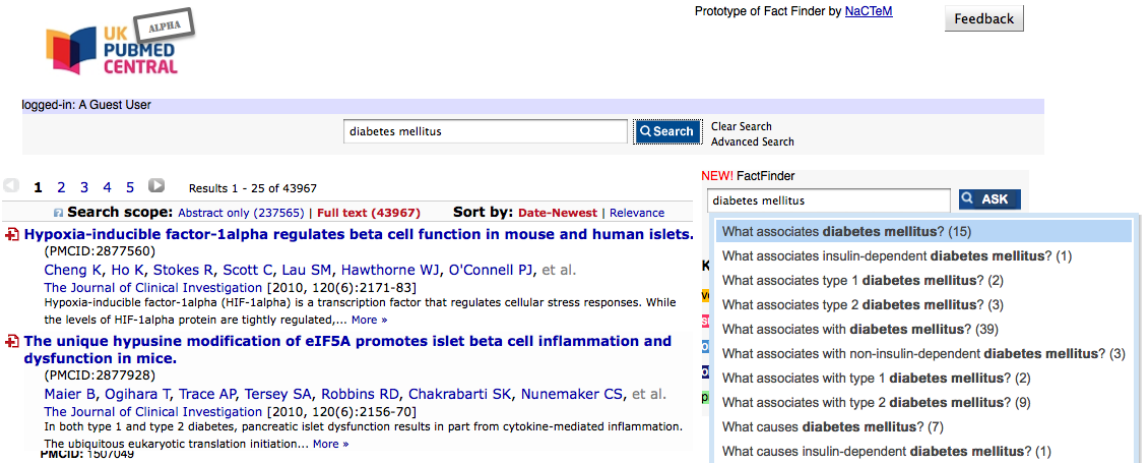

Figure 2: Illustration of the FactFinder prototype merged with the UKPMC Beta search engine