SkylineSearch: Semantic Ranking and Result Visualization for PubMed

Julia Stoyanovich\textsuperscript{2} Mayur Lodha\textsuperscript{1} William Mee\textsuperscript{1} Kenneth A. Ross\textsuperscript{1}

\texttt{jstoy@cis.upenn.edu \{mdl2130, wjm2107, kar\}@cs.columbia.edu}

\textsuperscript{1}CS Department, Columbia University, New York, NY, USA
\textsuperscript{2}CIS Department, University of Pennsylvania, Philadelphia, PA, USA

ABSTRACT

Life sciences researchers perform scientific literature search as part of their daily activities. Many such searches are executed against PubMed, a central repository of life sciences articles, and often return hundreds, or even thousands, of results, pointing to the need for data exploration tools. In this demonstration we present SkylineSearch, a semantic ranking and result visualization system designed specifically for PubMed, and available to the scientific community at skyline.cs.columbia.edu. Our system leverages semantic annotations of articles with terms from the MeSH controlled vocabulary, and presents results as a two-dimensional skyline, plotting relevance against publication date. We demonstrate that SkylineSearch supports a richer data exploration experience than does the search functionality of PubMed, allowing users to find relevant references more easily. We also show that SkylineSearch executes queries and presents results in interactive time.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—Scientific databases; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

General Terms

Algorithms, Human Factors, Performance

Keywords

Data Exploration, Skyline

1. INTRODUCTION

Literature search is a central task in scientific research. PubMed (www.pubmed.gov) is the most significant bibliographic source in life sciences, and many researchers, practitioners, and students search PubMed as part of their daily activities. PubMed currently indexes over 20 million articles that date back to 1865, and the number of new articles increases steadily from year to year.

PubMed articles are annotated by a staff of indexers with terms from the Medical Subject Headings (MeSH) controlled vocabulary (www.nlm.nih.gov/mesh). MeSH organizes terms into a hierarchy, allowing searching at various levels of specificity. MeSH annotations are currently used for query expansion: a query that matches a term will return all articles annotated with that term or with its descendants in MeSH. Due to the size of PubMed, and to the comprehensive query expansion strategy, many queries return hundreds, or even thousands, of results.

For example, the query Autoimmune Diseases AND Pregnancy returned over 11,000 results on October 28, 2010. The PubMed search interface supports the sorting of results by publication date, first or last author, title, and publication venue. Navigating a result set that contains 11,000 matches in sorted order, by a single metadata field, has several limitations. First, while publication date is a valuable dimension in scientific literature search, a user may want to find relevant articles even if they are not among the most recent few. Second, while author and journal names can help the user find a reference of which he is already aware, sorting on these metadata fields does not conveniently support true information discovery.

In [5] we proposed to use MeSH annotations for relevance ranking of PubMed search results. We found that MeSH has a unique structure, and we termed it a scoped polyhierarchy. Figure 2 presents a portion of MeSH that models autoimmune and connective tissue diseases. The hierarchy is a tree of nodes, with one or several nodes mapping to a single term. Thus, MeSH is a polyhierarchy. For example, the term “Arthritis, Rheumatoid” (RA) is represented by two nodes in MeSH. A term may span a subtree of different shape in different branches, and so the polyhierarchy is scoped. So, RA has six descendants in the “Rheumatic Diseases” subtree and four descendants in the “Autoimmune Diseases” subtree, representing that not all types of RA are autoimmune.

In [5], we developed several notions of relevance that are appropriate for scoped polyhierarchies, and showed how result relevance...
can be computed efficiently on the scale of PubMed and MeSH. We also presented a novel algorithm that efficiently computes a two-dimensional skyline of results, plotting relevance against publication date. In this demonstration we present SkylineSearch, an on-line system that implements and extends the techniques of [5], and integrates them into a complete result visualization system. A screenshot of our system is presented in Figure 1.

Our system improves on our prior work in two important ways. First, in [5] we showed that ranking and skyline computation that use our proposed measures of relevance can be executed efficiently, given a set of candidate matches. In this demonstration we will show that the end-to-end computation, described in Section 2, can also be executed in interactive time. Second, in our prior work we presented results of a user study that explored the effectiveness of our relevance measures for ranking. The SkylineSearch system goes further, and shows that our relevance measures can lead to a richer user experience when used to construct a two-dimensional skyline. We will demonstrate that SkylineSearch allows users to find relevant references more easily than does the PubMed search interface. Demonstration scenarios are outlined in Section 3.

2. SYSTEM OVERVIEW

2.1 User Interface

Figure 1 presents a screenshot of the SkylineSearch user interface. The user interacts with the system by issuing a keyword query to be executed against PubMed. A typical query will consist of one or several keywords, optionally connected with AND or OR, e.g., autoimmune diseases AND pregnancy and Alzheimer disease. Individual keywords, or groups of keywords, may also be designated as MeSH terms, for example connective tissue diseases[MeSH Terms] AND autoimmune diseases[MeSH Terms]. The user also chooses one of the similarity measures, described in Section 2.2, and sets the number of skyline contours, defined below.

SkylineSearch evaluates the query, computes the skyline of results, and displays the skyline, using the process described in Section 2.3. In addition to computing the skyline of the result set, our system computes up to 20 skyline contours. Skyline contours are useful for highlighting points that are close to the skyline, and that might be of interest to the user. A skyline contour is defined inductively as follows: (i) a point belongs to the $1^{st}$ contour iff it belongs to the skyline of the whole data set; (ii) a point belongs to the $k^{th}$ contour iff it belongs to the skyline of the data set obtained by removing points from the $1^{st}$ through $k-1^{st}$ contours.

Candidate results are processed by SkylineSearch in batches, and skyline points are presented to the user as soon as the first batch has been processed. The user may mouse over a point that represents a result, displaying a description of the corresponding PubMed article. The description lists the title, authors, publication venue and date, and MeSH annotations, and allows to navigate to the article in PubMed. The user may save and tag results for future reference.

2.2 Similarity Measures

At the heart of SkylineSearch lie the similarity measures used to access the relevance of a PubMed article to a query. In [5] we introduced three measures that exploit the unique structure of MeSH. We now briefly describe these measures, using a small scoped polyhierarchy in Figure 3. In defining the measures we use the notion of a term-scope. The term-scope of a query, denoted by termScope(Q), is the set of MeSH terms that the query matches, along with all descendants. The term-scope of a document, termScope(D), is defined analogously. In Figure 3, $Q = \{B, E\}$ and $D = \{D, G, F\}$; termScope(Q) = \{B, C, G, F, E\} is represented by a red circle, termScope(D) = \{D, G, F\} is represented by a blue circle.

SkylineSearch implements five similarity measures, illustrated in Figure 3. Coverage measures how exhaustively the document answers the query, and is computed as $\frac{|\text{termScope}(Q)\cap\text{termScope}(D)|}{|\text{termScope}(D)|}$. Specificity measures how specific the document is to the query, and is computed as $\frac{|\text{termScope}(Q)\cap\text{termScope}(D)|}{|\text{termScope}(Q)|}$. Jaccard is the harmonic mean of coverage and specificity. These are adaptations of the TermSim measure from [5]. Conditional counts the number of ancestor-descendant pairs $(s, t)$, where $s \in \text{termScope}(Q)$ and $t \in \text{termScope}(D)$; it implements CondSim from [5]. Balanced is similar to Conditional, but it normalizes the contribution of each query term to the score; it implements BalancedSim from [5].
Articles published over a year ago are shown in the higher-scoring part of the skyline, and include mostly survey articles (e.g., PubMed ids 16275205 and 15928277). These articles are not easily accessible via a sorted list provided by PubMed, since they do not start appearing until position 482 in the ranked list, and are not easy to tell apart from case reports and pharmacological studies.

With Conditional and Balanced similarity, results follow a similar trend, because scoping is not pronounced in this query. The actual set of skyline points is somewhat different than with Jaccard, because of different score normalization (see Section 2.2).

**Use Case 2:** Autoimmune Diseases AND Pregnancy Complications matches two MeSH terms, spanning subtrees of size 64 and 91, respectively. Similar trends are observed for this query with Jaccard similarity as for use case 1. Additionally, because both query terms exhibit scoped polyhierarchy features in their subtrees, and because they span large subtrees of different size — Conditional, and particularly Balanced, relevance measures produce very interesting results on the skyline. The skyline and early contours computed with the Balanced measure contain several articles about pregnancy complications in Systemic Lupus Erythematosus and in Antiphospholipid Syndrome, e.g., PubMed ids 19897518 and 17283586. These autoimmune disorders are commonly associated with pregnancy complications, and so these are high-quality answers. The answers appear prominently on the skyline, but would not have been easy to find in PubMed’s sorted list, because they are not among the most recent few results.

### 4. RELATED WORK

Rada and Bicknell [4] considered ranking MEDLINE documents using MeSH, and modeled the distance between the query and the document as the mean path-length between all pairs of document and query terms. This is one of several distance-based measures; see also Lee and Kim [3]. In contrast, we compare sets of terms via common descendants, not via common ancestors. We believe that descendant-based similarity is more appropriate, because query expansion in PubMed also incorporates descendants of a query term. Query expansion and ranking are parts of the same ranked retrieval process, and should agree on the semantics of relevance.

Several systems for bibliographic search in life sciences have been developed, see Kim and Rebholz-Schuhmann [2] for a review. The system closes ours, GoPubMed [1], uses three ontologies — GO, MeSH and Uniprot, to organize PubMed query results in a faceted hierarchy. When multiple ontology terms appear in the hierarchy, the system allows navigation by each of the terms. Unlike in our work, contributions of multiple terms are not reconciled into a single score, and no skyline visualization of results is provided.

### 5. REFERENCES


