ScienScan – an efficient visualization and browsing tool for academic search.

Daniil Mirylenka and Andrea Passerini

Department of Information Engineering and Computer Science
University of Trento, Via Sommarive 5, 38123, Trento, Italy
{dmirylenka,passerini}@disi.unitn.it

Abstract. In this paper we present ScienScan[1] – a browsing and visualization tool for academic search. The tool operates in real time by post-processing the query results returned by an academic search engine. ScienScan discovers topics in the search results and summarizes them in the form of a concise hierarchical topic map. The produced topical summary informatively represents the results in a visual way and provides an additional filtering control. We demonstrate the operation of ScienScan deploying it on top of the search API of Microsoft Academic Search.

1 Introduction

We often use academic search engines for exploratory tasks, such as reviewing related work or investigating unfamiliar topics, when the goal is not to retrieve specific publications but rather to improve our understanding of the topic in question. Performing exploratory search, we tightly interact with the search engine, refining our queries based on the retrieved results. Presented as endless unstructured lists, typical search results are difficult to examine and interpret, often making our exploratory search task tedious and even frustrating.

In this paper we describe ScienScan – an academic search tool that provides concise visual summaries of the query results. These summaries convey useful information about the topical structure of the result set in an intuitive way, without the user having to sift through individual items. In addition, the produced summaries serve as a filtering control, allowing the user to focus on the relevant subtopics of the query, and thus find papers more efficiently.

ScienScan presents a novel approach to visualization and browsing of academic search results. It employs state-of-the-art external tools and services as well as newly developed methods, and can run on top of third-party search engines. Based on the practical solutions, ScienScan is, to the best of our knowledge, the only available prototype tool providing this type of functionality.

2 Web Interface

ScienScan is a Web tool with an interface of a typical search engine (see Figure 2). Users type queries into the search box and obtain the list of search results. In

[1] http://scienscan.disi.unitn.it/

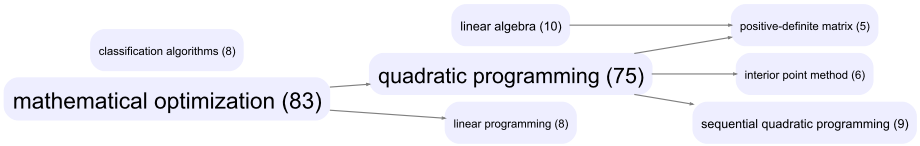

Fig. 1. Example summary of 100 search results for the query "quadratic programming".

addition to the standard controls, ScienScan displays a topic map of the retrieved results. The topic map is a small taxonomy of topics built so as to summarize the search results in the most informative way (see Figure 1 for an example).

The topic map is a directed acyclic graph, in which the child nodes represent the subtopics of the parent nodes. The nodes in the topic map represent topics relevant to the search results. Each topic covers a subset of the results in such a way that a parent topic always covers all the results covered by its child topics and, possibly, some additional results. The number of covered results is displayed in the parentheses near the topic title, with the font size of the title being proportional to this number. When the user clicks on a topic in the hierarchy, the topic and its subtopics become highlighted, and the displayed results get restricted to those covered by the selected topic.

The user can control the number of nodes in the topic map by moving the slider. ScienScan builds multiple instances of the topic maps of various sizes, and moving the slider switches between these instances, making the displayed topic map grow or shrink visually. The algorithms of ScienScan are implemented in such a way that bigger topic maps are built incrementally from smaller ones. This makes the computations efficient, and ensures that no topic disappears from the map when the map size is increased.

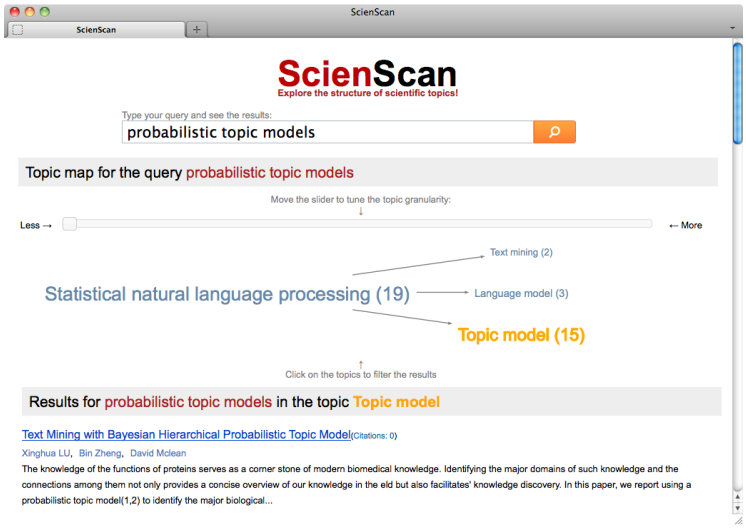

Fig. 2. The interface of `ScienScan`.

3 Architecture

When the user submits the search query, ScienScan performs the following steps:

1. forward the query to an existing search engine and collect the results;
2. link the results to Wikipedia articles with the help of a topic annotator;
3. build the topic graph based on the retrieved Wikipedia articles, their categories and hierarchical relations between them;
4. summarize the constructed topic graph into a concise topic map;
5. visualize the topic map using a graph-drawing tool.

A detailed description of the steps 2-5 can be found in [4].

Search engine. The current version of ScienScan relies on the search API of `Microsoft Academic Search`. The API restricts usage to 200 requests per minute and returns up to 100 search results per request. The latter restriction is not too limiting, as users seldom look beyond the first one hundred results.

Topic annotation. During this step the search result titles and abstracts are mapped to Wikipedia articles, which can be viewed as fine-grained meaningfully labeled topics. Recently, there have been proposed multiple methods and tools for annotating texts with links to Wikipedia articles. Some of the implementations have publicly available demos, restricted web service APIs, and standalone tools requiring a snapshot of Wikipedia database (see, for example, TAGME, Machine Linking, DBPediaSpotlight and Wikipedia Miner[3]). We deployed an instance of the Wikipedia Miner web service for the purpose of topic annotation.

Building the topic graph. The following steps expand the set of article topics into a large topic graph based on the network of Wikipedia categories:

1. retrieve the parent categories of the discovered articles (transforming the set of articles into a bipartite article-category graph);
2. connect the categories according to their taxonomic relations in Wikipedia (transforming the bipartite graph into a general directed topic graph);
3. merge similar topics and break the cycles (making the topic graph acyclic);
4. detect and extend the main topic of the query (that with the most search results, making the topic graph more detailed in the area of the main topic).

Summarizing and displaying the topic graph. At this step we have to reduce the graph containing about three hundred nodes to a concise taxonomy, such as shown in Figure 1. A good taxonomy must possess a number of important properties, such as high coverage of the search results, relevance, high frequency and low redundancy of the included topics. The current version of ScienScan applies a frequency-based heuristic algorithm to select the most informative set of nodes from the topic graph. A new version being under development uses a more advanced summarization algorithm based on structured-output prediction [4]. The mentioned algorithms prescribe which topics from the original topic graph should be included into the summary. In order to completely define the summary, we connect the topics with the minimum number of links that still maintain the hierarchical relations induced by the original graph. After the topic map is built, we submit it to the `graphviz` package [1] for visualization. The `dot` algorithm used in `graphviz` for drawing directed graphs produces the layered layout appropriate for displaying topic hierarchies.

4 Related work and discussion

To the best of our knowledge, there exist no other online academic search tools providing structured visual representations of the query results. Current popular scholarly services include publishers' digital libraries (such as ACM or IEEE), search engines (Google Scholar, Microsoft Academic Search or CiteSeer) and social networking sites (Mendeley, CiteULike, ResearchGate). These services typically provide browsing based on metadata, such as publication venue, authors or year, or a predefined topic categorization scheme. As categorization schemes are independent of the current query and rather coarse-grained, no visualization of the search results is provided based on them, nor on the metadata attributes.

In contrast to available tools, in the literature there have been proposed numerous sophisticated methods for detecting and visualizing research topics. The main approaches to this problem include frequent keyword-based methods, analyses of citation graph, and probabilistic topic models, the latter probably representing the most developed class of methods (see [2] for an example). In the context of search result visualization, these methods have the following shortcomings: *a*) they typically require access to the whole corpus of papers rather than only current results, *b*) (except for keyword-based methods) they do not provide short meaningful labels for discovered topics. Keyword-based methods have an additional shortcoming in that the topics correspond to verbatim keywords. ScienScan avoids these shortcomings by relying on Wikipedia-based topics.

`TAG MY SEARCH` [5] is an example of Wikipedia-based topic discovery applied to a related task of *general Web search* result clustering. Unlike ScienScan, `TAG MY SEARCH` uses only articles but not categories of Wikipedia to represent topics, and thus performs flat rather then hierarchical grouping of the search results. Action Science Explorer and Sci^2 represent publication collections as networks (for instance, citation-based), and provide visualization and exploration tools typical for network analysis, such as clustering and filtering based on metadata and network statistics. In contrast, ScienScan builds a higher-level view that is focused on the explicit labeled semantic topics and their hierarchical relations.

Acknowledgments. This research was partially supported by grant PRIN 2009LNP494 (Statistical Relational Learning: Algorithms and Applications) from Italian Ministry of University and Research.

References

1. E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software – Practice & Experience*, 30(11), 2000.
2. Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles. Detecting topic evolution in scientific literature: how can citations help? In *CIKM*, pages 957–966. ACM, 2009.
3. D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194(0):222 – 239, 2013.
4. D. Mirylenka and A. Passerini. Learning to grow structured visual summaries for document collections. In *ICML Workshop on Structured Learning*, 2013.
5. U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita. Topical clustering of search results. In *WSDM*, pages 223–232. ACM, 2012.