3-30-2016

Web-based Visual Interfaces Designed for Searching on Collections of Research Papers

Christopher Bellis
Indiana University - Purdue University Fort Wayne

Chris Perry
Indiana University - Purdue University Fort Wayne

Max Fowler
Indiana University - Purdue University Fort Wayne

Follow this and additional works at: http://opus.ipfw.edu/stu_symp2016

Recommended Citation
http://opus.ipfw.edu/stu_symp2016/28

This Book is brought to you for free and open access by the IPFW Student Research and Creative Endeavor Symposium at Opus: Research & Creativity at IPFW. It has been accepted for inclusion in 2016 IPFW Student Research and Creative Endeavor Symposium by an authorized administrator of Opus: Research & Creativity at IPFW. For more information, please contact admin@lib.ipfw.edu.
Web-based Visual Interfaces Designed for Searching on Collections of Research Papers

Christopher Bellis, Chris Perry
Max Fowler, Beomjin Kim (kimb@ipfw.edu)
Department of Computer Science
Indiana University-Purdue University Fort Wayne

Visual Interface

Our system is broken into two tiers. The visualization uses a force graph algorithm, in which a given paper (Tier 1) or term (Tier 2) is tethered to the queries it is related to. This tethering is given a strength, which determines the layout of the graph. Papers are scored by relevance of queries within them.

Index Construction

Our system uses Apache's Lucene to construct indices over text content. PDFs were scraped from Google Research. The PDFs were then converted into plain text, then passed through Lucene. The index for a document keeps track of word count and sentence positions, while removing common words (*"stop" words), punctuation, and other special symbols.

Project Overview

When attempting to conduct research on a given subject, finding relevant sources can prove a notable challenge. Current tools, such as online journal databases using a keyword-based search are often lacking, and provide you with a plethora of trivial, only tangentially related papers. Further, these searches also make it difficult to find papers that cross into different domains.

This poster presents a two tier visual interface for searching collections of documents. Tier 1 presents the results of a three query search. Tier 2 is designed to not only show the contents of the documents, but to aid in search term refinement for the first tier. Both tiers are designed to be applied to an arbitrary digital library of unstructured text content. An index is constructed over this content using Lucene.

We aim to both make finding the highest quality source in a set of documents easier and refining searches to be more targeted to a specific area of interest a trivial task rather than a difficult thought process.

Design Methodology

Our system was designed as a series of code modules: one for indexing the digital library, one for providing a connection to the library, one for displaying the visualization, and one so that the code to be reusable for future research projects. For example, the visualization to the left side of the poster could be applied to crime scene data, rather than journal papers. This would allow for correlating and categorizing criminal reports.

For the sake of development, our digital library is a collection of journal papers scraped from Google Research. The index is constructed using Lucene, a popular, open source search package.

This work is intended to explore aspects of visual analytics that have not yet been explored. Prior digital library work has focused predominately on the use of meta-data only or on the use of the content of structured data, such as books. Research papers provided an untapped niche to explore.

Current Progress

Tier 1 provides an intuitive way to find the best paper from a set of papers for a given set of queries. The visualization is easy to navigate and the black and white color scale much more readily shows paper quality than traditional, text based searching systems. For example, a Google search would give no indication of journal paper quality.

Tier 2 allows the user to see, at a glance, if a query is featured often in a paper by node size. Further, the presence of document content lets a user painlessly swap in a new term into their searches, to aid search refinement.

Our system provides ease of searching. ease of selecting good sources, and ease of search construction while using a visual interface that is more pleasing and requires less work to navigate than a text based search engine.

Update Plan

Usability tests will be conducted in late March and early April, in order to judge the effectiveness of our approach. These tests will be taken by first year Computer Science students, such that they have some domain knowledge but not so much as to develop a bias beforehand. The test will compare our visual interface to traditional, text based searching.

Future updates include:
1. Including meta-data, such as conference quality and publication year, into Tier 1.
2. Including a firmer cut off for related terms, so only the best picks make it onto the Tier 2 graph. This will reduce on some of the clutter.
3. Increased support for multi-term queries -- the current system supports them, but it can lead to unusual behavior in the Tier 1 layout and Tier 2 results.

Our work will be featured at the Modern Artificial Intelligence and Cognitive Science Conference at the University of Dayton, Ohio, on April 22-23, 2016, in the paper Exploring Web-based Visual Interfaces for Searching Research Articles on Digital Library Systems, by Fowler et al.