Web Semantics: Science, Services and Agents on the World Wide Web 8 (2010) 169–175

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Invited paper

Visualizing search results and document collections using topic maps

David Newman[a,*,1], Timothy Baldwin[a,b], Lawrence Cavedon[a], Eric Huang[a], Sarvnaz Karimi[a], David Martinez[a], Falk Scholer[a], Justin Zobel[a,b]

[a] NICTA Victorian Research Laboratory, Melbourne, Australia
[b] Department of Computer Science and Software Engineering, University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 2 June 2009
Accepted 26 March 2010
Available online 20 April 2010

Keywords:
Topic modelling
Visualizing document collections
Clustering
Dimensionality reduction

ABSTRACT

This paper explores visualizations of document collections, which we call *topic maps*. Our topic maps are based on a topic model of the document collection, where the topic model is used to determine the semantic content of each document. Using two collections of search results, we show how topic maps reveal the semantic structure of a collection and visually communicate the diversity of content in the collection. We describe techniques for assessing the validity and accuracy of topic maps, and discuss the challenge of producing useful two-dimensional maps of documents.

1. Introduction

While advances have been made in semantically characterizing web documents, less progress has been made in creating meaningful and useful semantic visualizations of text document collections. The need for visualizing document collections arises in many situations, particularly for users wanting to gain a better understanding of an entire set of search results. While many users are focused on finding specific information, there are large numbers of users that want to find and understand all information about a particular topic, and understand the span (both depth and breadth) of their search results. An accurate and intuitive visualization of search results could facilitate this understanding.

For example, a medical researcher may want to systematically review treatments for spinal cord injuries, determine current best practices, and identify controversial interventions. In this situation, it is critical to exhaustively find all relevant information, so researchers often have to manually scan and digest large numbers of search results from broader and less specific queries. Another example might be an NSF program manager trying to understand a large collection of research proposals on rapid climate change. The program manager could benefit from a visual map of all the proposals to better understand the relationships between the various lines of research.

We explore how topic maps – visual displays of document collections – can help with these types of problems. Our topic maps are created by first learning a topic model of a text document collection. Topic models (which can be viewed as the Bayesian version of latent semantic analysis) are useful for extracting semantic content from collections of documents. After topic modelling, we project onto two dimensions the topic representation of documents to create the topic map visualization.

In this paper we present examples of topic maps and use these examples to describe validation techniques. Validation is important since there is no correct or unique answer for what makes a useful topic map. We start by motivating topic mapping with a description of our topic mapping tool in Section 2. We then step back and examine the component steps of topic mapping. We first show the relevance and validity of topic modelling in Section 3. Next we compare three projection methods for making topic maps in Section 4. Finally, we conclude the paper with a discussion in Section 5.

2. NICTA topic mapping tool

As part of the Elsevier Grand Challenge, NICTA developed a topic mapping tool to address the goals of "improving the interpretation and identification of meaning in articles relating to the life

* Corresponding author.
E-mail addresses: newman@uci.edu (D. Newman), tim@ccse.unimelb.edu.au (T. Baldwin), lawrence.cavedon@nicta.com.au (L. Cavedon), eric.huang@nicta.com.au (E. Huang), sarvnaz.karimi@nicta.com.au (S. Karimi), david.martinez@nicta.com.au (D. Martinez), falk.scholer@rmit.edu.au (F. Scholer), jz@ccse.unimelb.edu.au (J. Zobel).
[1] On leave from the Department of Computer Science, University of California, Irvine, United States.

1570-8268/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2010.03.005

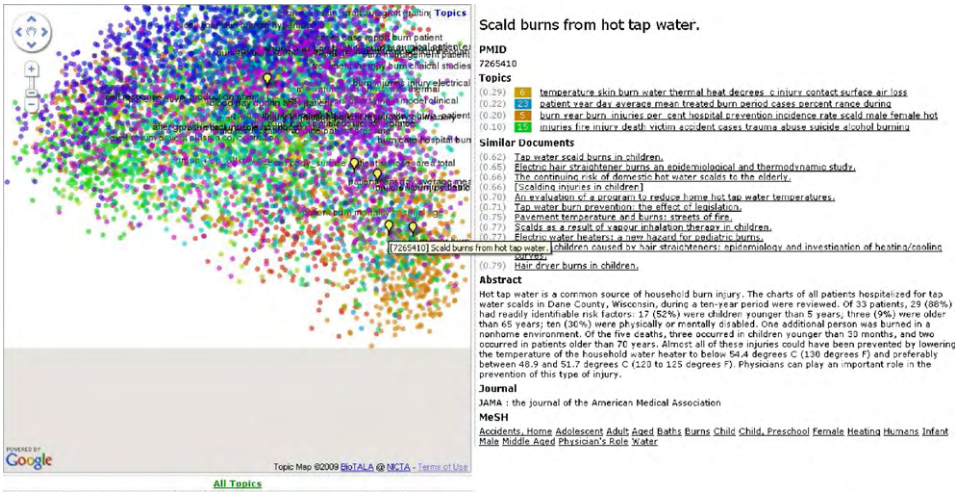

Fig. 1. Screen shot of NICTA topic mapping tool, showing 10,172 PubMed search results from a query about *burns*. The left side displays a Google-maps topic map visualization of all 10,172 search results. Each dot is a search result, color coded by the article's primary topic. The right side provides detail about one particular search result ("scald burns from hot tap water"), showing topics, similar documents and the abstract.

sciences" and "interpreting, visualizing and connecting knowledge more effectively."[2]

Our topic mapping tool takes as input a collection of text documents (that may correspond to a set of search results from some query). A topic model is learned for the collection, producing a set of topics that describe the collection, and multiple topic assignments to each document in the collection. Using the topic coordinates of each document, we project the set of documents onto two dimensions with the goal of preserving nearest neighbors (i.e. similar documents should appear close together on the topic map). This two-dimensional map is then rendered at various resolutions and cut into image tiles which are accessed using pan and zoom via the Google maps api.

A screen shot of the NICTA topic mapping tool is shown in Fig. 1. This screen shot shows 10,172 PubMed search results from a query about *burns*. The left side displays the topic map visualization of all 10,172 search results, color coded by each article's primary topic. The right side provides detail about one particular search result ("scald burns from hot tap water"), showing the component topics, similar documents and the abstract.

In this tool one can navigate and browse the collection of search results, both by clicking around the map on the left, and by navigating via text links on the right. The two sides of the display are coordinated—selecting a document on the map will bring it up on the right, and visa versa. One can toggle the display of individual topics to indicate the spatial extent of that topic.

While this topic mapping approach is not novel (e.g. see Ref. [6]), in this paper we experiment with different mapping approaches and describe various validation techniques. For validation purposes, we used two databases of text documents: full text articles provided by Elsevier for the Elsevier Grand Challenge, and MEDLINE abstracts accessed from PubMed. We created focused collections that were produced by issuing search queries against these two databases. We issued the Boolean query "bayesian" against the Elsevier database (restricting to articles from life-science journals), which returned 1230 full-length articles. We refer to this collection as the *Bayesian* search results. We also issued the query "acute spinal cord injury" to PubMed, which returned 4169 records. Of those records, 3558 included abstracts, and this made up our collection of *Spinal Cord* search results. These two document collections were then turned into the standard bag-of-words representation for modelling.

3. Topic modelling

Topic models (also known as Latent Dirichlet Allocation or LDA models) are probabilistic models for document collections, and are seen by many in the machine learning community as the state-of-the-art for extracting semantic information from collections of text documents [1,4]. A topic model learns a set of thematic topics from words that tend to occur together in documents. In the topic model, an integer ID t is assigned to every word in every document according to $P(topic = t) \propto P(word|t)P(t|doc)$, where $t \in 1 \ldots T$, and T is the specified number of topics to learn. After an initial random assignment of topics to words, the Gibbs sampler iteratively updates these topic assignments until the topics, $P(word|topic)$, and topic mixtures, $P(topic|doc)$, converge. The set of topics is a semantic basis for representing the entire collection, and a useful way to represent individual documents.

A single topic is a multinomial distribution over words, $P(word|topic)$, where the probability mass is concentrated on a small fraction of words that relate to the topic. For example, Fig. 2 shows the distribution of the top nine words in a topic relating to rat models of pain learned from *Spinal Cord* search results.

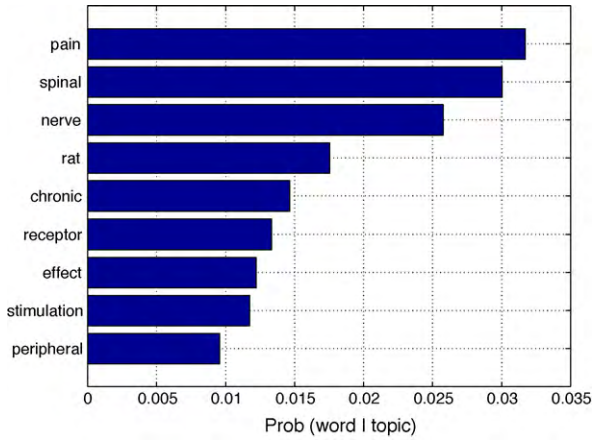

Fig. 2. A topic is a focused multinomial distribution over words. This histogram shows one topic from the $T = 20$ topic model of *Spinal Cord* search results.

[2] http://www.elseviergrandchallenge.com.