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.

© 2010 Elsevier B.V. All rights reserved.

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.

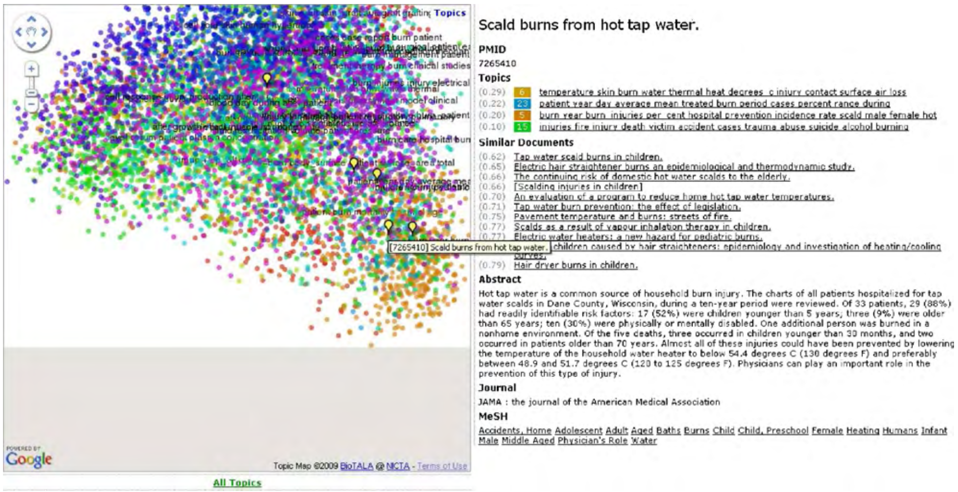
2. NICTA topic mapping tool

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


Topics
[7265410] Scald burns from hot tap water.
POWERED BY Google
Topic Map ©2009 BioTALA @ NICTA - Terms of Use
All Topics

Scald burns from hot tap water.

PMID

7265410

Topics

(0.29) 6 temperature skin burn water thermal heat degrees c injury contact surface air loss
(0.22) 23 patient year day average mean treated burn period cases percent range during
(0.20) 5 burn year burn injuries per cent hospital prevention incidence rate scald male female hot
(0.10) 15 injuries fire injury death victim accident cases trauma abuse suicide alcohol burning

Similar Documents

(0.62) Tap water scald burns in children.
(0.65) Electric hair straightener burns an epidemiological and thermodynamic study.
(0.66) The continuing risk of domestic hot water scalds to the elderly.
(0.66) [Scalding injuries in children]
(0.70) An evaluation of a program to reduce home hot tap water temperatures.
(0.71) Tap water burn prevention: the effect of legislation.
(0.75) Pavement temperature and burns: streets of fire.
(0.77) Scalds as a result of vapour inhalation therapy in children.
(0.77) Electric water heaters: a new hazard for pediatric burns.
children caused by hair straighteners: epidemiology and investigation of heating/cooling curves.
(0.79) Hair dryer burns in children.

Abstract

Hot tap water is a common source of household burn injury. The charts of all patients hospitalized for tap water scalds in Dane County, Wisconsin, during a ten-year period were reviewed. Of 33 patients, 29 (88%) had readily identifiable risk factors: 17 (52%) were children younger than 5 years; three (9%) were older than 65 years; ten (30%) were physically or mentally disabled. One additional person was burned in a nonhome environment. Of the five deaths, three occurred in children younger than 30 months, and two occurred in patients older than 70 years. Almost all of these injuries could have been prevented by lowering the temperature of the household water heater to below 54.4 degrees C (130 degrees F) and preferably between 48.9 and 51.7 degrees C (120 to 125 degrees F). Physicians can play an important role in the prevention of this type of injury.

Journal

JAMA : the journal of the American Medical Association

MeSH

Accidents, Home Adolescent Adult Aged Baths Burns Child Child, Preschool Female Heating Humans Infant Male Middle Aged Physician's Role Water

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.

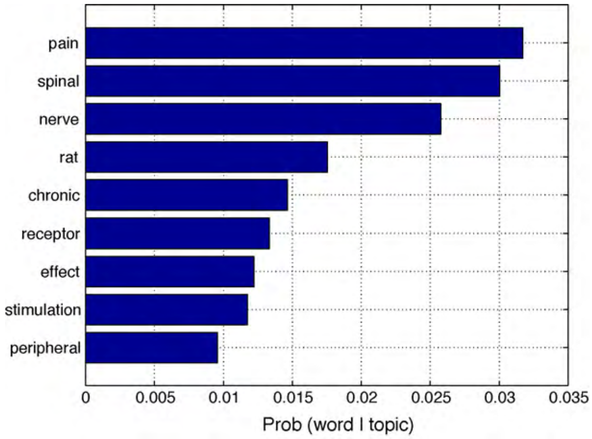


Fig. 2. A topic is a focused multinomial distribution over words. This histogram shows one topic from the $T = 20$ topic model of *Spinal Cord* search results.

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

Table 1
Learned topics ($T = 20$) for *Bayesian* search results. The number in parentheses is percent of all words assigned to that topic. The topic ID, e.g. $t12$, is arbitrary.

$t12$. (5%) model variables data spatial species . . .
$t17$. (4%) model system process control dynamic . . .
$t19$. (4%) water energy period temperature concentration . . .

$t14$. (5%) system information user data decision . . .
$t16$. (5%) pattern structure fact general complex . . .

$t1$. (6%) protein genes gene sequence sequences . . .
$t4$. (3%) test infection animal prevalence species . . .
$t13$. (4%) population genetic individual marker association . . .

$t3$. (5%) effect treatment risk trial group . . .
$t6$. (5%) patient risk disease clinical coronary . . .
$t8$. (4%) cell effect dose response concentration . . .
$t18$. (5%) subject brain region activity activation . . .

$t2$. (5%) robot set learning object action . . .
$t5$. (6%) speech model word speaker recognition . . .
$t7$. (6%) algorithm target problem measurement estimation . . .
$t9$. (5%) image cluster algorithm images wavelet . . .
$t10$. (5%) signal noise component source frequency . . .
$t11$. (5%) network algorithm input training output . . .
$t15$. (7%) model parameter distribution data prior . . .
$t20$. (6%) data test set class classification . . .

Topic modelling provides an instant summary of the entire contents of a collection. Tables 1 and 2 show the topics learned for *Bayesian* and *Spinal Cord* search results, grouped by related topics. In *Bayesian* we see a wide range of topics—from speech recognition ($t5$, assigned to 6% of all words), to coronary heart disease risk ($t6$, assigned to 5% of all words). Note that topics are not a mutually exclusive partitioning of terms in the vocabulary, for example *test* occurs in different contexts in topics $t4$ and $t20$. For *Spinal Cord*, we see three topics relating to rat models ($t1$, $t9$ and $t4$), and other topics that have distinct foci, e.g. imaging ($t7$), bladder infection ($t10$) and rehabilitation ($t3$). Here we do not label the topics, but when labels are needed they can be suggested by domain experts. While labels are useful in many end-user applications, the list of words in the topic is the most accurate and transparent way of communicating a topic's meaning.

At the individual document level, topics provide a useful low-dimensional semantic representation. For example, the article "Cellular mechanisms of hyperalgesia and spontaneous pain in a spinalized rat model of peripheral neuropathy: changes in myelinated afferent inputs implicated." has $P(t4|doc) = 0.66$ (pain spinal nerve rat . . .) and $P(t1|doc) = 0.09$ (cord spinal injury . . .) as its two biggest topics. Similar documents would likely include one or more of topics $t4$ and $t1$. The probabilistic framework is very flexible—for example, using Bayes rule, we can also find the most likely documents in topic $t4$ by computing $P(doc|t4)$.

Table 2
Learned topics ($T = 10$) for *Spinal Cord* search results.

$t7$. (14%) patient spinal cervical cord spine acute imaging . . .
$t8$. (8%) sci muscle acute subject during after bone . . .
$t10$. (9%) patient acute complication bladder infection . . .

$t1$. (9%) cord spinal injury after lesion rat axon . . .
$t2$. (8%) treatment injury cord spinal acute effect after . . .
$t3$. (13%) patient injury sci rehabilitation spinal outcome year . . .
$t5$. (10%) spinal cord injury acute injuries management review . . .
$t9$. (11%) cord spinal group injury after rat +/−. . .

$t4$. (9%) pain spinal nerve rat chronic receptor effect . . .
$t6$. (10%) cell injury expression sci system protein nervous . . .

3.1. Validation of topics learned by the topic model

Since the topic model is an unsupervised learning method, we want to demonstrate that the set of learned topics is stable, and not an artifact of one particular run of the topic model. A simple test is to learn multiple independent topic models, each starting with a different random initialization of topic assignments to words. Below we list the most likely words in five different learned topics relating to coronary heart disease, from five independent topic models of *Bayesian*. The similarity in the top-8 words suggests that this topic is consistently learned from the data:

- patient risk disease clinical coronary exercise cardiac heart . . .
- patient risk disease clinical coronary exercise cardiac test . . .
- patient risk clinical disease coronary exercise cardiac heart . . .
- patient risk clinical treatment trial disease group coronary . . .
- patient risk trial clinical disease treatment coronary exercise . . .

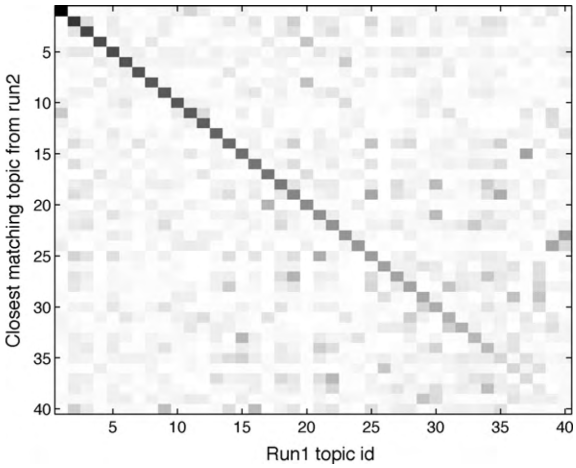


Fig. 3. Distance of topics in run1 to reordered topics in run2 ($T = 40$ topic model of *Bayesian* search results).

Fig. 3 shows the repeatability of an entire set of topics from two independent runs of the topic model learning $T = 40$ topics. The image shows the distance between each topic in run1 to each topic in run2, reordered so that the closest matching topics fall on the diagonal. Again, we see a high degree of stability over independent topic model runs.

Do learned topics for our *Spinal Cord* search results match up with MeSH headings (the medical subject headings provided by MEDLINE)? Table 3 shows the learned topics most associated with selected MeSH headings prevalent in the *Acute SCI* search results. We determine the closest matching topic to each MeSH heading by computing $P(x, y)/P(x)P(y)$, where x is the event that a document includes a given topic and y is the event that a document includes the mesh heading. A higher number means more depen-

Table 3
Validation of learned topics ($T = 40$) against MeSH headings, using specific mutual information. The number in parentheses measures $P(x, y)/P(x)P(y)$, where x is the event that a document includes the topic and y is the event that a document includes the mesh heading. A higher value means higher dependence between the two events.

MeSH	Topic
Pain	(12.7) $t10$. pain nerve neuropathic dorsal rat peripheral mechanical model . . .
Spinal fractures	(8.3) $t36$. cervical spine fracture injuries patient injury trauma . . .
Paraplegia	(3.9) $t18$. aortic paraplegia thoracic artery ischemia repair aorta . . .
Spinal cord compression	(5.9) $t3$. spinal compression case epidural report acute . . .
Magnetic resonance imaging	(7.2) $t27$. imaging MRI magnetic resonance cord patient . . .

Table 4
Comparison of different unsupervised learning methods showing general agreement in selected learned topics (*Bayesian* search results using $T = 20$).

LDA	Patient risk trial clinical disease treatment coronary exercise cardiac event heart evidence . . .
k-Means	Patient risk clinical group test disease data cardiac heart coronary exercise myocardial event . . .
NMF	Patient risk trial treatment clinical disease test coronary exercise cardiac heart level . . .
LDA	Speech model word speaker recognition system adaptation acoustic hmm feature performance . . .
k-Means	Speech model speaker word system noise data recognition acoustic performance adaptation signal hmm . . .
NMF	Model speech speaker word system acoustic hmm recognition feature features performance training . . .
LDA	Protein genes gene sequence sequences species expression DNA phylogenetic cell binding tree . . .
k-Means	Protein genes gene data sequences sequence species population genetic DNA region level expression . . .
NMF	Protein genes gene cell expression sequence data gene-expression DNA microarray RNA level . . .

dence between the two events. Here we see good correspondence between the learned topics and the MeSH headings. We also see the finer resolution of topics at $T = 40$, learning topics specifically relating to MRI ($t27$) and Fracture ($t36$).

3.2. Comparison with other unsupervised learning methods

How do topics learned by the topic model compare to semantic concepts learned by other unsupervised learning methods? Using the *Bayesian* search results, we learned a set of $T = 20$ topics using three fundamentally different models: the topic model (LDA); *k*-Means clustering [3]; and nonnegative matrix factorization (NMF) [8], which is a modified version of singular value decomposition (i.e. latent semantic analysis) devised for positive data.

Table 4 shows a comparison of selected topics learned by these three models. We observe that each method independently finds very similar topics in three selected areas (coronary heart disease, speech recognition and gene sequencing), and the list of most relevant words in each case is very similar. While not shown, the similarity across methods extends throughout the entire set of topics.

While this provides further evidence that learned topics are real, accurate and stable, we suggest that topic modelling (LDA) is the preferred approach for extracting semantic topics. The topic model represents documents as a mixture of topics, which is a more flexible representation than *k*-Means clustering, where a document belongs to just one topic. This richer representation means that one can navigate to related articles via any of the topic facets in a given article (this is not possible for *k*-Means). Furthermore, since the topic model is based on a generative probabilistic model, it can be flexibly extended to include other metadata and attributes, such as authors, citation links, and subject headings, and therefore it is also preferred over NMF [5].

4. Projecting topic-modelled documents onto two dimensions

Having established that learning a topic model is an accurate and reliable way of determining the semantic content of documents, how best can we show this information in a two-dimensional (2D) display? We are faced with another dimensionality reduction problem: after reducing the document dimension from 10,000s of words to 10s of topics, we need to further reduce the document dimension from 10s of topics down to 2 dimensions.

For this projection task, we can either work directly with the $D \times T$ matrix of topic coordinates for each document, or the $D \times D$ matrix of pairwise inter-document distances, as measured by the L_1 distance between each document's topic distribution. The distribution of these pairwise distances (Fig. 4, computed for *Spinal Cord*) points to a potential difficulty in producing a faithful 2D projection. This distribution of all pairwise distances (which has a similar shape to the distribution of $D - 1$ distances from any one document to all other documents) tells us that while there are a small number of documents that are close (i.e. similar) to a given document, the vast majority of documents appear almost equally far away at a distance of about 1.6 units in L_1 space. So in 2D, from any given document, nearly all the documents should appear in an approximate circle or band, with the given document at the center. Clearly this geometry is impossible to achieve for all documents simultaneously. Therefore we expect to have significant approximation error when projecting the set of D documents onto two dimensions. This particular problem of high-dimensional data has been discussed by Weber et al. [10].

With this challenge in mind, we compare three different algorithms for producing two-dimensional topic maps of a set of D documents: principal component analysis (PCA), local linear embedding (LLE) and force directed layout (FDL). Since a faithful layout is unlikely, we aim to optimize the 2D arrangement of documents with the following desiderata: (i) if two documents are topically close, then they should appear close on the map; and (ii) if two documents are close on the map, then they should be topically close.

PCA finds the 2D projection that maximizes the variance of the data. Note that PCA is essentially the same as linear multidimensional scaling, a well-known method for producing locations from similarity data. LLE is designed to more accurately represent local structure of the data (which is aligned with our desideratum of topically similar documents appearing close on the map), by doing a reconstruction of k nearest neighbors [9]. Finally, FDL is a technique for graph layout techniques typically used for non-fully connected graphs. FDL is based on attractive spring forces (between

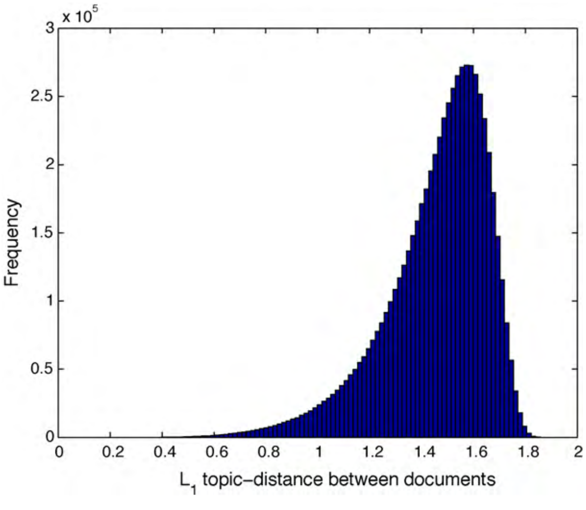


Fig. 4. Distribution of L_1 distances between all pairs of documents. From any one document, there are a small number of close documents, but all other documents appear to be equally far away (at a distance of 1.6 units).

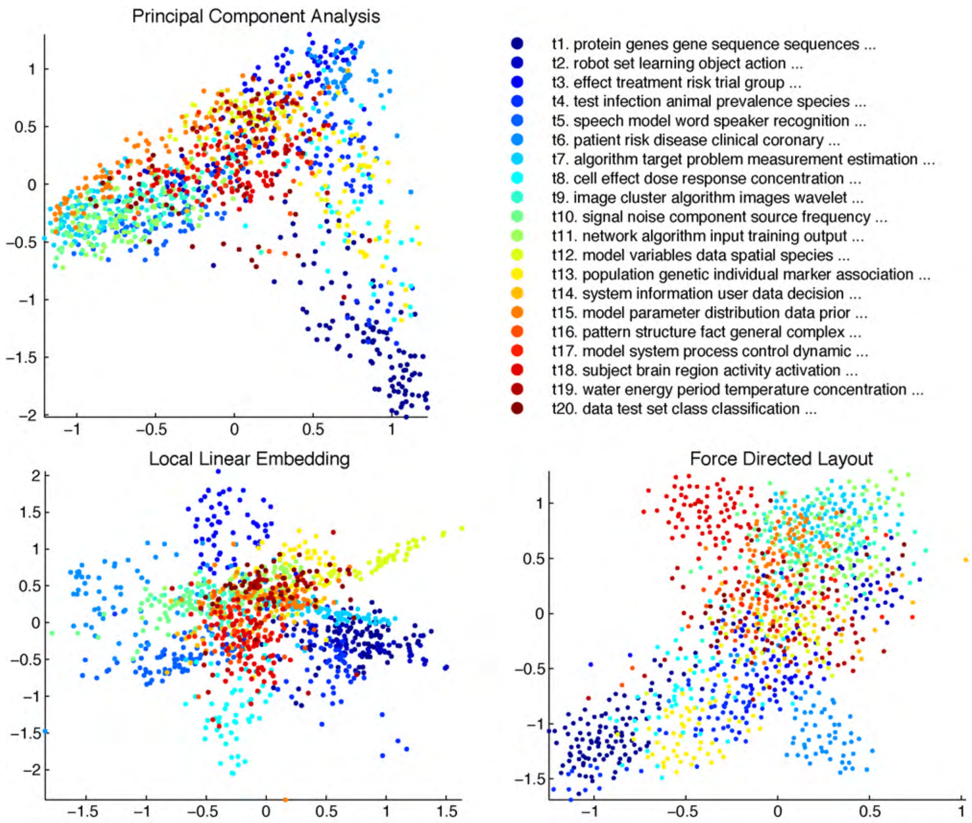


Fig. 5. PCA, LLE and FDL topic maps for 1230 *Bayes* search results.

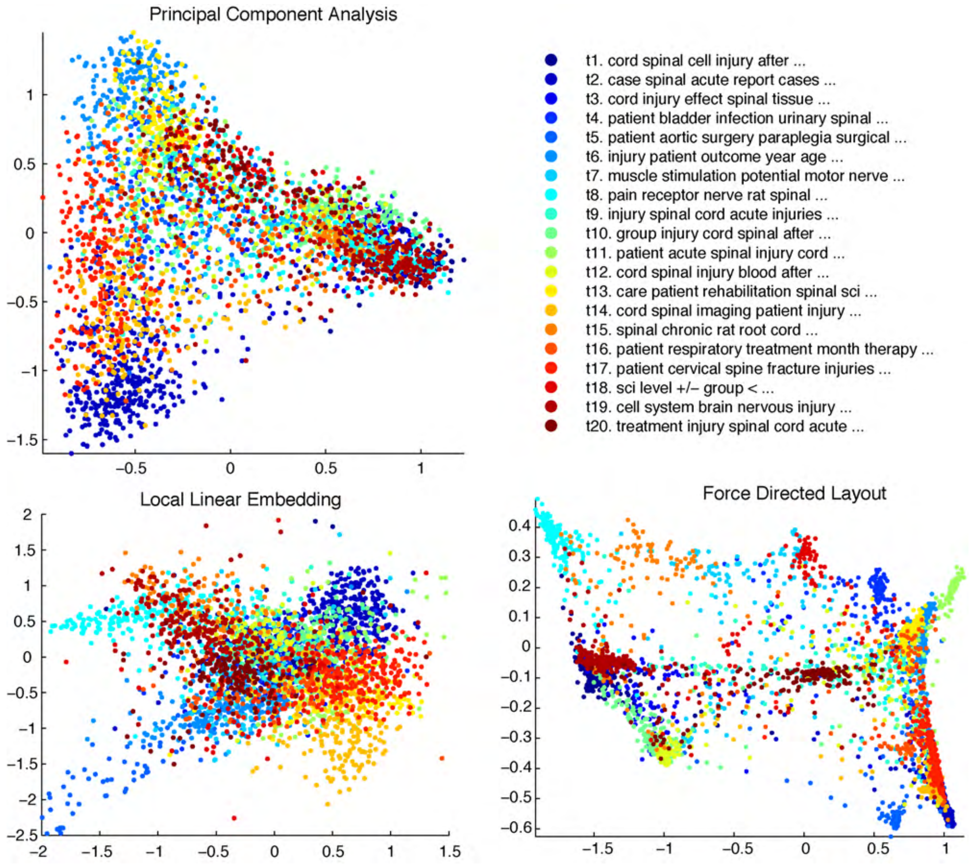


Fig. 6. PCA, LLE and FDL topic maps for 3558 *Spinal Cord* search results.

close documents) and repulsive forces (in high-density regions) [2]. We chose these three methods to produce a variety of layouts for further examination.

4.1. Topic maps

Topic maps for *Bayes* search results are shown in Fig. 5 for PCA, LLE and FDL projections. Each document is represented by a colored dot. The majority topic in the document determines the color of the dot (this is the simplest color coding, since a document is actually a mixture of topics), and the topic colors are shown in the legend. The scales on the maps are arbitrary, and we include them to use in diagnostics. In all three topic maps we observe large-scale coherence in the spatial arrangement of documents on a particular topic (i.e. local blobs of color for the twenty topics). But comparing the three projections shows that there is no consistent geometry to the large scale arrangement of these topic blobs. For example "*t*18. subject brain region . . ." is relatively central in PCA and LLE, but off to the side in FDL. The topic maps show topics that co-occur, for example "*t*15. model parameter . . ." appears with "*t*20. data test set . . ." in all three topic maps.

The *Spinal Cord* topic maps are shown in Fig. 6. Here we used a $T = 20$ topic run to produce and color the topic maps. From these two sets of topic maps we can make some observations. The PCA projection tends to make a more convex distribution of documents than LLE, which tends to have a radial or spoke-like pattern. These two projections tend to produce distributions of documents with relatively constant density, in contrast to the FDL layout, which produces (for *Acute SCI*) a more clumped/sparse arrangement of documents. This variable density is possibly more an artifact than actually present in the data.

After reviewing these topic maps, there is nothing in particular that suggests that one projection method is better than another. Some of the projections seem to separate the topics better, but does this produce an overall more useful topic map? Given the fundamentally different nature of objective functions for PCA, LLE and FDL, and the inherent difficulty of faithfully representing this data in 2D, it is not surprising that the three topic maps look different for each collection. We continue by proposing a metric to compare topic maps.

4.2. Evaluation of projections

Which of these three projection methods produces the best (and most useful) topic map? The answer to this question depends on one's objective judgement (and perhaps user evaluation, which is beyond the scope of this paper). In each method, a different objective function is optimized, but which objective is most aligned with our desideratum of preserving closeness? A simple spot check of local accuracy in Fig. 7 shows detail from the PCA topic map in Fig. 6 around the document "Frequency and spectrum of pulmonary diseases in patients with chronic renal failure associated with spinal cord injury." The figure shows the 40 topically closest documents to this document (numbered 1 . . . 40). The lack of numbers in this closeup view shows that the PCA projection has difficulty in preserving the nearest neighbors to this document.

This figure suggests one approach for measuring the accuracy of a topic map. We want to measure how often documents are close on the map, but unrelated, and how often documents are topically close, but far apart on the map. To do this we compare topic-based pairwise document distances with their corresponding pairwise distances on the map. But since we are more interested in preserving closeness, we focus our attention on short distances (i.e. close documents), both in the higher-dimensional topic space, and on the two-dimensional topic map. Consider the shortest 10% of all D^2 pairwise distances on the topic map. This is same as considering all documents within a particular radius from any given document on the map. Fig. 8 (top) shows a box/whisker plot of *precision* —the distribution of L_1 topic distances corresponding to the shortest 10%. This is analogous to precision in IR in that it is measuring the probability that a "returned" document (i.e. within a fixed radius on the map) is relevant. Fig. 8 (bottom) shows a box/whisker plot of *recall* —how many relevant (close) documents appear on the topic map within the fixed radius. This panel shows the distribution of Euclidean map distances corresponding to the shortest 10% of all D^2 topic distances. These are standard box/whisker plots that show

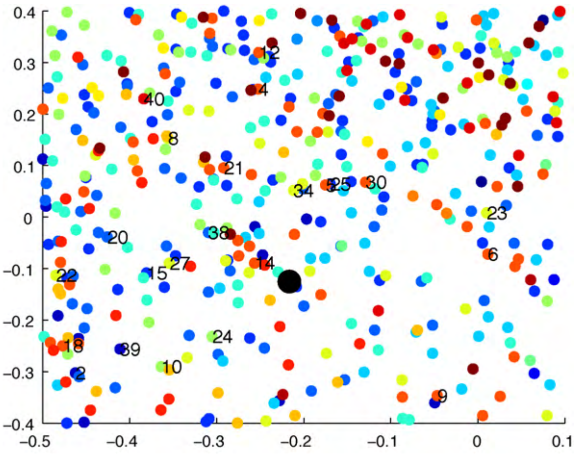


Fig. 7. Forty nearest neighbors of PMID 6828807 (larger black dot located near center) numbered 1 . . . 40, taken from PCA projection of *Spinal Cord* search results.

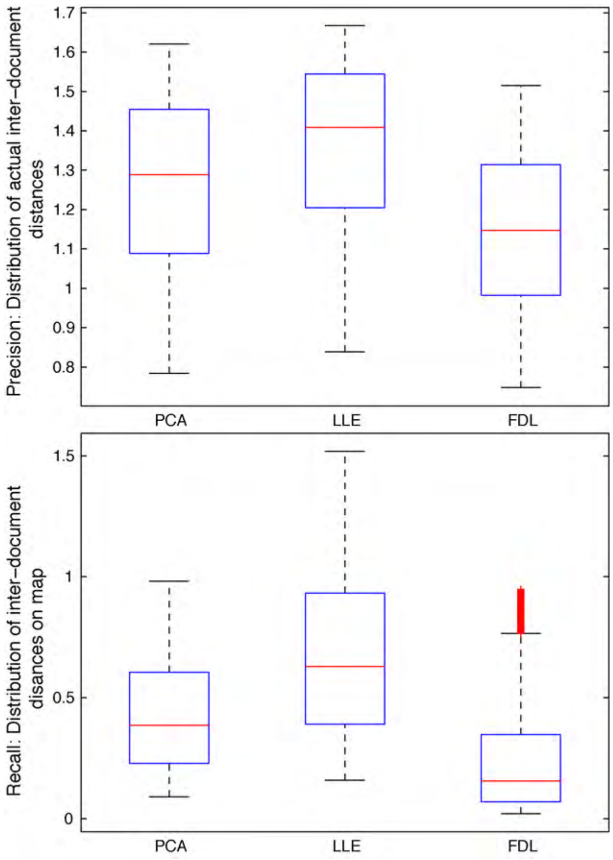


Fig. 8. Distribution of map distances that correspond to the 10% shortest topic-based distances (top). Distribution of topic distances that correspond to the 10% shortest map distances (bottom). Results are for *Spinal Cord*.

the 25th percentile, median and 75th percentile with a box and 5th and 95th percentile with the whiskers. With these *precision* and *recall* statistics, lower numbers are better (i.e. the best projection preserves closeness).

From these plots we may conclude – for this dataset – that the FDL projection is the best and the LLE projection is the worst. For a constant density of points this would be true. However, we observe that the FDL projection's density of dots is highly variable. The clumpiness of the FDL topic map could artificially improve *recall* by virtue of the tight bunching of dots on the map. Surprisingly LLE, which is meant to preserve local structure (and thereby be better at preserving closeness), performs worse than PCA using these measures. In fact all three projection methods have relatively low precision, compared to the theoretical optimum.

We conclude this section by commenting that evaluating topic maps is a complicated task. The precision and recall box/whisker plots do convey some useful metrics to compare methods, but a more thorough comparison should use multiple metrics, including the precision and recall measures we have shown.

5. Discussion and conclusions

In this paper we have examined the steps for making topic maps, and discussed techniques for evaluating accuracy. We separately evaluate the two steps involved in making topic maps—topic modelling, and projection onto two dimensions. We show that topic modelling has relatively high accuracy, and the topics learned for a collection provide a useful basis to understand and browse that collection. However, it is more difficult to assess the two-dimensional projection of a set of topic-modelled documents since there is no single measure that determines accuracy or validity. The topic maps do visually convey some information about a document collection, but perhaps the ultimate test is human assessment which evaluates performance of a particular task using the topic map.

Other researchers have produced topic maps using a variety of methods. A large scale semantic map of Wikipedia was created based on links between pages [7] and force directed layout was used to create large-scale topic maps of NIH funding [6].

While not explored in this paper, we have the opportunity to use many visualization techniques to improve the usability and interpretability of our topic maps, beyond simple color coding by a single topic. We often have other attributes or metadata (such as authors, citation links and subject headings) that can enhance topic maps, using any combination of color, shape, size and text annotations.

One conclusion from this work is that local topic maps (which dynamically show a few dozen closely related documents) may be more accurate. While topics are a useful way to organize an entire collection, producing a static global topic map of the collection may have limited value for exploring the collection. Therefore local topic maps may ultimately be more useful for better understanding and navigating local structure in a collection.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work builds on earlier work referenced in [6], and we gratefully acknowledge the ideas developed there. We also acknowledge the contribution of discussions with The Global Evidence Mapping Initiative (GEM: http://www.evidencemap.org). We thank GEM for helping us understand the query process used in systematic reviews, and for supplying us with sample queries.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[2] K.W. Boyack, R. Klavans, K. Börner, Mapping the backbone of science, Scientometrics 64 (3) (2005) 351–374.
[3] I.S. Dhillon, D.S. Modha, Concept decompositions for large sparse text data using clustering, Machine Learning 42 (1/2) (2001) 143–175.
[4] T. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004).
[5] T.L. Griffiths, M. Steyvers, J.B.T. Tenenbaum, Topics in semantic representation, Psychological Review 114 (2) (2007) 211–244.
[6] B. Herr, E. Talley, G. Burns, D. Newman, G. LaRowe, The NIH visual browser: an interactive visualization of biomedical research, in: 13th International Conference on Information Visualization, 2009.
[7] T. Holloway, M. Bozicevic, K. Börner, Analyzing and visualizing the semantic coverage of wikipedia and its authors, Complexity 12 (3) (2007) 30–40.
[8] D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, 401.
[9] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[10] R. Weber, H.-J. Schek, S. Blott, A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces, in: VLDB'98: Proceedings of the 24rd International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.