Evaluating Topic Representations for Exploring Document Collections

Nikolaos Aletras (corresponding author)
Computer Science
University College London
nikos.aletras@gmail.com

Timothy Baldwin
Computing and Information Systems
The University of Melbourne
tb@ldwin.net

Jey Han Lau
Department of Philosophy
King's College, London
jeyhan.lau@gmail.com

Mark Stevenson
Computer Science
University of Sheffield
mark.stevenson@sheffield.ac.uk

Abstract

Topic models have been shown to be a useful way of representing the content of large document collections, for example via visualisation interfaces (topic browsers). These systems enable users to explore collections by way of latent topics. A standard way to represent a topic is using a term list, i.e. the top-n words with highest conditional probability within the topic. Other topic representations, such as textual and image labels, have also been proposed. However there has been no comparison of these alternative representations. In this paper, we compare three different topic representations in a document retrieval task. Participants were asked to retrieve relevant documents based on pre-defined queries within a fixed time limit, presenting topics in one of the following modalities: (1) lists of terms, (2) textual phrase labels, and (3) image labels. Results show that textual labels are easier for users to interpret than term lists and image labels. Moreover, the precision of retrieved documents for textual and image labels is comparable to the precision achieved by representing topics using term lists, demonstrating that labelling methods are an effective alternative topic representation.

1 Introduction

In recent years, a large amount of information has been made available on-line in digital libraries, collections and archives. Much of this information is stored in unstructured format (such as text) and is not organised using any classification system. The sheer volume of available information can be overwhelming for users, making it very difficult to find specific information or even explore such collections. The majority of search interfaces rely on keyword-based search. However, this approach only works when users have sufficient domain knowledge to be able to generate appropriate queries but this is not always the case. Users may not know what information is available or not be sufficiently familiar with the information to be able to select appropriate keywords.

There are, of course, alternatives to keyword-based search which are useful in situations where the user is not familiar with the collection. Approaches that provide the user with an overview of the information available in the collection have proved useful for information seeking tasks such as exploratory search (Marchionini, 2006) and sense-making (Hearst, 2009). For example, faceted browsing has proved useful for exploratory search (Collins et al., 2009; Hearst, 2006; Smith et al., 2006). However, these approaches often presuppose a consistent classification scheme for the collection. Unfortunately these do not exist for all collections (e.g. because the collection is constructed from a disparate set of documents with no classification scheme, or is aggregated across collections with incompatible schemes) and manual classification is impractical for all but the smallest of collections.

These problems can be ameliorated by using large-scale automatic data-analysis techniques to present the unstructured information to the user in a distilled manner which they can browse through. Topic models (Blei et al., 2003; Hofmann, 1999) offer an unsupervised, data-driven means of capturing the themes discussed within document collections. These are represented via a set of latent variables called topics. Each topic is a probability distribution over words occurring in the collection such that words that co-occur frequently are each assigned high probability in a given topic. Topic models also represent documents in the collection as probability distributions over the topics that are discussed in them.

Topic models have been shown to be a useful way of representing the content of large document collections, for example via visualisation interfaces (topic browsers) (Chaney and Blei, 2012; Ganguly et al., 2013; Gretarsson et al., 2012; Hinneburg et al., 2012; Snyder et al., 2013). These systems enable users to navigate through the collection by presenting them with sets of topics. Topic models are well suited for use in these interfaces since they are able to identify underlying themes in collections and can be applied at low human cost, through the use of unsupervised learning.

Topics are often represented using a list of terms, i.e. the top-n words with highest marginal probability within a topic, such as *school, student, university, college, teacher, class, education, learn, high, program*. Alternative representations, such as textual phrase labels (e.g. EDUCATION for our example topic), can potentially assist with the interpretations of topics, and researchers have developed methods to generate these automatically (Mei et al., 2007; Lau et al., 2010; Lau et al., 2011). Approaches that make use of alternative modalities, such as images (Aletras and Stevenson, 2013b), have also been proposed, with the advantage that they are language independent and potentially provide at-a-glance access to the collection.

Intuitively, labels represent topics in a more accessible manner than the standard term list approach. However, there has not, to our knowledge, been any empirical validation of this intuition, a shortcoming that this paper aims to address, in carrying out a task-based evaluation of different topic model representations. In this, we compare three approaches to representing topics: (1) a standard term list, (2) textual phrase labelling, and (3) image labelling. These are used to represent topics generated from a digital archive of news-wire stories, and evaluated in an exploratory search task.

The aim of this study is to compare different topic representations within a document retrieval task. We aim to understand the impact of different topic representation modalities in finding relevant documents for a given query, and also measure the level of difficulty in interpreting the same topics through different representation modalities. We are interested in answering the following research questions:

1. which topic representations are suitable within a document browser interface?

2. what is the impact of different topic representations on human search effectiveness for a given query?

Section 2 reviews previous work on automatically labelling topics and the use of topic models to create search interfaces. Section 3 introduces an experiment in which three approaches to topic labelling are applied and evaluated within an exploratory search interface. The results of the experiment on exploratory search are presented in Section 4, followed by intrinsic evaluation of the labels generated by the different methods in Section 5.

2 Related Work

In early research on topic modelling, topics were represented as ranked lists of terms with the highest probability, and textual labels were sometimes manually assigned to topics for convenience of presentation of research results (Mei and Zhai, 2005; Teh et al., 2006).

The first attempt to automatically assigning labels to topics is described by Mei et al. (2007). In their approach, a set of candidate labels is extracted from a reference collection using noun chunks and bigrams with high lexical association. Then, a relevance scoring function is defined which minimises the distance between the word distribution in a topic and the word distribution in candidate labels. Candidate labels are ranked according to their relevance, and the top-ranked label is chosen to represent the topic.

Magatti et al. (2009) introduced an approach for labelling topics that relies on two manually labelled hierarchical knowledge resources: the Google Directory and the OpenOffice English Thesaurus. The Automatic Labelling Of Topics algorithm computes the similarity between LDA-inferred topics and categories in the topic tree, a pre-existing hierarchical set of labelled categories, by computing scores using six standard similarity measures. The label for the most similar category in the topic tree is assigned to the LDA topic.

Lau et al. (2010) proposed selecting the most representative term from a topic as its label by computing the similarity between each word and all others in the topic. Several sources of information are used to identify the best label, including pointwise mutual information scores, WordNet hypernymy relations and distributional similarity. These features are combined in a re-ranking model.

Lau et al. (2011) proposed a method for automatically labelling topics, using Wikipedia article titles as candidate labels. A set of candidate labels is generated in four phases. Primary candidate labels are generated from Wikipedia article titles by querying using topic terms. Then, secondary labels are generated by chunk parsing the primary candidates to identify chunk n-grams that exist as Wikipedia article titles. Outlier labels are identified using a word similarity measure (Grieser et al., 2011) and removed. Finally, the top-5 topic terms are added to the candidate set. The candidate labels are ranked using information from word association measures, lexical features and an information retrieval technique.

Mao et al. (2012) introduced a method for labelling hierarchical topics which makes use of sibling and parent–child relations of topics. Candidate labels are generated using a similar approach to the one used by Mei et al. (2007). Each candidate label is then assigned a score by creating a distribution based on the words it contains, and measuring the Jensen-Shannon divergence between this and a reference corpus. Results show that incorporating information about the relations between topics improves label quality.

Hulpus et al. (2013) use the structured data in DBpedia[1] to label topics. Their approach maps topic words to DBpedia concepts and identifies the best ones using graph centrality measures, assuming that words co-occurring in text likely refer to concepts that are closer in the DBpedia graph.

Cano Basave et al. (2014) presented a method for labelling LDA topics trained on social media streams, i.e Twitter, using summarisation techniques. Their method generates labels which exist in the Twitter stream rather than relying on external knowledge sources.

Aletras and Stevenson (2014) introduced an unsupervised graph-based method that selects textual phrase labels for topics. PageRank (Page et al., 1999) is used to weigh the words in the graph and score the candidate labels.

In contrast, Aletras and Stevenson (2013b) proposed a method for labelling topics using images rather than text. A set of candidate images for a topic is retrieved by querying an image search engine with the top-n topic terms. The most suitable image is selected using PageRank. The ranking algorithm makes use of textual information from the metadata associated with each image, as well as visual features extracted from the analysis of the images themselves.

Topic modelling has been used to support browsing in large document collections (Gardner et al., 2010; Newman et al., 2010; Wei et al., 2010; Chaney and Blei, 2012; Hinneburg et al., 2012; Chuang et al., 2012; Ganguly et al., 2013; Snyder et al., 2013). The collection is often presented to users as a set of topics. Users can access documents in the collection by selecting topics of interest. The vast majority of topic-based browsers developed so far have relied on using lists of terms to represent the topics, and have not made use of the previous research on automatically generating labels for topics. We address this limitation by making use of three approaches to labelling topics within a topic-based browser and carrying out experiments to compare their effectiveness.

Reuters Topic Category (Query)	No. Docs.
Travel & Tourism	314
Domestic Politics (USA)	27,236
War - Civil War	16,615
Biographies, Personalities, People	2,601
Defence	4,224
Crime, Law Enforcement	10,673
Religion	1,477
Disasters & Accidents	3,161
International Relations	19,273
Science & Technology	1,042
Employment/Labour	2,796
Government Finance	17,904
Weather	1,190
Elections	5,866
Environment & Natural World	1,933
Arts, Culture, Entertainment	1,450
Health	1,567
European Commission Institutions	1,046
Sports	18,913
Welfare, Social Services	775

Table 1: Number of documents in each Reuters Corpus topic category

3 Methodology

We conducted an experiment to compare three topic representations: (1) lists of terms, (2) textual phrase labels, and (3) image labels. Users were provided with an interface representing a set of topic models derived from a collection and asked to search for

[1] http://dbpedia.org

Modality	Label
Term list	*report, investigation, officials, information, intelligence, former, government, documents, alleged, fbi*
Textual Phrase Label	*Federal Bureau of Investigation*
Image Label	

Table 2: Labels generated for an example topic.

documents that were relevant to a set of queries.

We chose to use a search task given the widely used and well understood methodologies that are available. Interfaces based on topic models are more suited to document browsing but quantifying performance is less straightforward for this task.

3.1 Document Collection

We make use of a subset of the Reuters Corpus (Rose et al., 2002), which is both freely available and has manually-assigned topic categories associated with each document. The topic categories are used both as queries in the retrieval task and to provide relevance judgements to determine the accuracy of the documents retrieved by users.

20 topic categories were selected and 100,000 documents randomly extracted from the Reuters Corpus. Each document is pre-processed by tokenisation, removal of stop words, and removal of words appearing fewer than 10 times in the collection, resulting in a vocabulary of 58,162 unique tokens. Table 1 shows the Reuters Corpus topic categories used to form the collection, together with the number of associated documents.

3.2 Topic Modelling

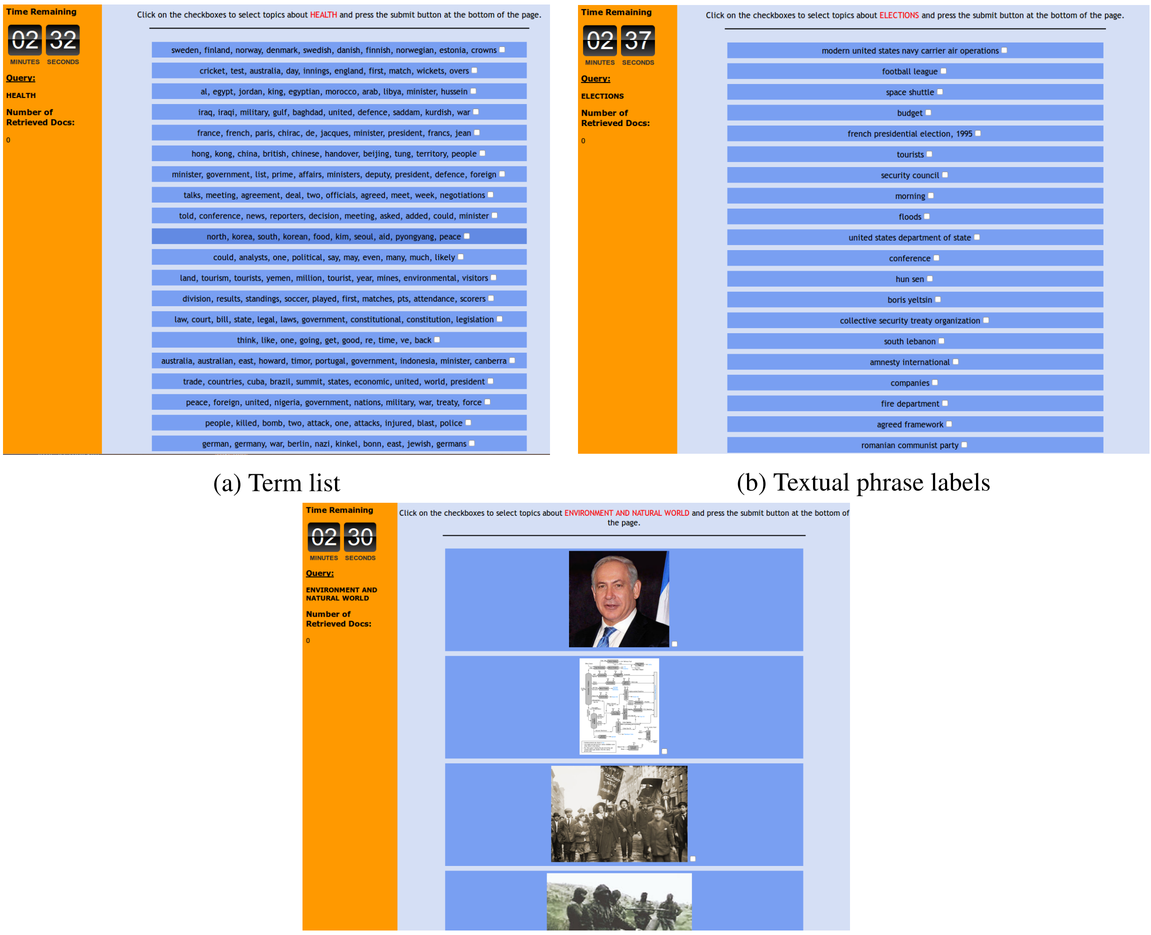
An LDA model was trained[2] over the document collection using variational inference (Blei and Jordan, 2003). The number of topics learned was set to $T = 100$ since topic interpretability in LDA tends to stabilise when $T \geq 100$ (Stevens et al., 2012). Default settings are used for all other parameters. Topics that are difficult to interpret were identified using the method of Aletras and Stevenson (2013a) and removed, leaving a total of 84 topics.

3.3 Topic Browsing Systems

The topic browsing system developed for this study is based on the publicly available Topic Model Visualisation Engine (TMVE) (Chaney and Blei, 2012). TMVE uses a document collection and an LDA model trained over that collection (see Section 3.2). It generates a topic browsing system with three main components: (1) a main page, (2) topic pages, and (3) document pages. The main page contains the list of automatically-generated topics. Each topic page shows a list of documents with the highest conditional probability given that topic. Document pages show the content of a document together with its topic distribution.

We created three separate browsing systems based on TMVE. The only difference between the three systems is the way in which they represent topics, namely: (1) term lists, (2) textual phrase labels, and (3) images. The term lists are created using a standard approach (see Section 3.3.1), the textual phrase labels are generated from Wikipedia article titles (Lau et al., 2011) (see Section 3.3.2), while the image labels are generated using publicly available images from Wikipedia (Aletras and Stevenson, 2013b) (see Section 3.3.3). By default, TMVE only

[2] We make use of the implementation provided by David Blei https://www.cs.princeton.edu/~blei/lda-c/index.html



Time Remaining

02 32

MINUTES SECONDS

Query:

HEALTH

Number of Retrieved Docs:

0

Click on the checkboxes to select topics about HEALTH and press the submit button at the bottom of the page.

- sweden, finland, norway, denmark, swedish, danish, finnish, norwegian, estonia, crowns □
- cricket, test, australia, day, innings, england, first, match, wickets, overs □
- al, egypt, jordan, king, egyptian, morocco, arab, libya, minister, hussein □
- iraq, iraqi, military, gulf, baghdad, united, defence, saddam, kurdish, war □
- france, french, paris, chirac, de, jacques, minister, president, francs, jean □
- hong, kong, china, british, chinese, handover, beijing, tung, territory, people □
- minister, government, list, prime, affairs, ministers, deputy, president, defence, foreign □
- talks, meeting, agreement, deal, two, officials, agreed, meet, week, negotiations □
- told, conference, news, reporters, decision, meeting, asked, added, could, minister □
- north, korea, south, korean, food, kim, seoul, aid, pyongyang, peace □
- could, analysts, one, political, say, may, even, many, much, likely □
- land, tourism, tourists, yemen, million, tourist, year, mines, environmental, visitors □
- division, results, standings, soccer, played, first, matches, pts, attendance, scorers □
- law, court, bill, state, legal, laws, government, constitutional, constitution, legislation □
- think, like, one, going, get, good, re, time, ve, back □
- australia, australian, east, howard, timor, portugal, government, indonesia, minister, canberra □
- trade, countries, cuba, brazil, summit, states, economic, united, world, president □
- peace, foreign, united, nigeria, government, nations, military, war, treaty, force □
- people, killed, bomb, two, attack, one, attacks, injured, blast, police □
- german, germany, war, berlin, nazi, kinkel, bonn, east, jewish, germans □

(a) Term list

Time Remaining

02 37

MINUTES SECONDS

Query:

ELECTIONS

Number of Retrieved Docs:

0

Click on the checkboxes to select topics about ELECTIONS and press the submit button at the bottom of the page.

- modern united states navy carrier air operations □
- football league □
- space shuttle □
- budget □
- french presidential election, 1995 □
- tourists □
- security council □
- morning □
- floods □
- united states department of state □
- conference □
- hun sen □
- boris yeltsin □
- collective security treaty organization □
- south lebanon □
- amnesty international □
- companies □
- fire department □
- agreed framework □
- romanian communist party □

(b) Textual phrase labels

Time Remaining

02 30

MINUTES SECONDS

Query:

ENVIRONMENT AND NATURAL WORLD

Number of Retrieved Docs:

0

Click on the checkboxes to select topics about ENVIRONMENT AND NATURAL WORLD and press the submit button at the bottom of the page.

(c) Image labels

Figure 1: Topic browsing interfaces.

supports the term list representation of topics, and required modification to support textual phrase and image labels. Table 2 shows examples of the labels generated by the three approaches for a sample topic.[3] In addition, in the topic page, each topic is associated with its top-300 highest-likelihood documents given the topic. We restrict the number of documents shown to the user for each topic to avoid the task becoming overwhelming.

3.3.1 Term lists

Term lists are generated using the default approach of TMVE, i.e. selecting the top-10 terms with the highest conditional probability within the topic. This is the standard approach to representing topics used within the topic modelling research community.

3.3.2 Textual Phrase Labels

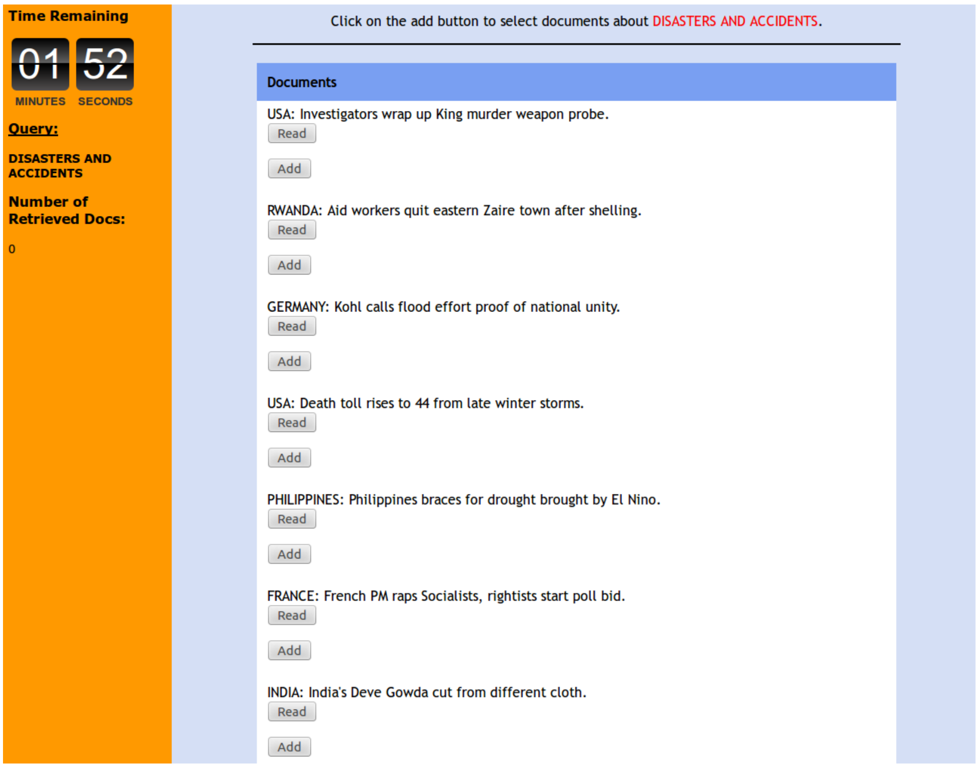
Textual phrase labels are generated using the approach of Lau et al. (2011), in two phases: candidate generation and candidate ranking.

In candidate generation, we use the top-7 topic terms[4] to search Wikipedia using Wikipedia's native search API and Google's site-restricted search. We collect the top-8 article titles returned from each of the search engines;[5] these constitute the primary

[3] Note that the textual phrase and image labels are created automatically (see Sections 3.3.2 and 3.3.3) and may contain errors. In this example the logo of the FBI may have been a more suitable image label than the one that was generated.

[4] From preliminary experiments we found that using the top-10 terms for search occasionally yields no results for a number of topics.

[5] The version of the Google search API used in the original paper limited the maximum number of results per query to 8.



Click on the add button to select documents about DISASTERS AND ACCIDENTS.

USA: Investigators wrap up King murder weapon probe.

RWANDA: Aid workers quit eastern Zaire town after shelling.

GERMANY: Kohl calls flood effort proof of national unity.

USA: Death toll rises to 44 from late winter storms.

PHILIPPINES: Philippines braces for drought brought by El Nino.

FRANCE: French PM raps Socialists, rightists start poll bid.

INDIA: India's Deve Gowda cut from different cloth.

Figure 2: Topic browsing: List of documents.

candidates. To generate more candidates, we chunk-parse the primary candidates to extract noun chunks and generate component n-grams from the noun chunks, excluding n-grams that do not themselves exist as Wikipedia titles. As this procedure generates a number of labels, we introduce an additional filter to remove labels that have low association with other labels, based on the RACO lexical association method (Grieser et al., 2011). The component n-grams that pass the RACO filter constitute the secondary candidates. Lastly, we include the top-5 topic terms as additional candidates.

In the candidate ranking phase, we generate a number of lexical association features of the label candidate with the top-10 topic terms: pointwise mutual information (PMI), Student's t-test, Pearson's χ^2 test, log likelihood ratio and two conditional probability variants. Term co-occurrence frequencies for computing these measures are sampled from the full collection of English Wikipedia with a sliding window of length 20 words. We also include two features based on the lexical composition of the label candidate: the raw number of terms it contains, and the proportion of terms in the label candidate that are top-10 topic terms. We combine all the features using a support vector regression model to rank the candidates.[6] The highest ranked candidate is selected as the textual phrase label for the topic.

3.3.3 Image Labels

We associate topics with image labels using the approach described by Aletras and Stevenson (2013b). We generate candidate labels using images from Wikipedia, available under the Creative Commons licence. The top-5 terms from a topic are used to query Bing using its Search API.[7] The search is restricted to English Wikipedia[8] with image search enabled. The top-20 images retrieved for each search are used as candidates for the topic, and are represented by textual and visual features.

Textual features are extracted from the metadata associated with the images. The textual information

[6] The model is trained using the labelled data collected by the authors in Lau et al. (2011).

[7] `http://datamarket.azure.com/dataset/bing/search`

[8] `http://en.wikipedia.org`

is formed by concatenating the *title* and the *url* fields of the search result. These represent, respectively, the web page title containing the image, and the image file name. The textual information is preprocessed by tokenisation and removal of stop words.

Visual information is extracted using low-level image keypoint descriptors, i.e. SIFT features (Lowe, 1999; Lowe, 2004) sensitive to colour information. Image features are extracted using dense sampling and described using Opponent colour SIFT descriptors provided by the *colordescriptor* package.[9] The SIFT features are clustered to form a visual codebook of 1,000 visual words using k-means clustering, such that each feature is mapped to a visual word. Each image is represented as a bag-of-visual words (BOVW).

A graph is created using the candidate images as the set of nodes. Edges between images are weighted by computing the cosine similarity of their BOVWs. Then, Personalised PageRank (PPR) (Haveliwala et al., 2003) is used to rank the candidate images. The personalisation vector of PPR is initialised by measuring average word association between topic words and image metadata based on PMI, as in Aletras and Stevenson (2013b). The image with the highest PageRank score is selected as the topic label.

3.4 Task

The aim of the task was to identify as many documents relevant to a set of queries as possible. Each participant had to retrieve documents for 20 queries (see Table 1), with 3 minutes allocated for each query. In addition to the query (e.g. *Travel & Tourism*), participants were also provided with a short description of documents that would be considered for the query (e.g. *News articles related to the travel and tourism industries, including articles about tourist destinations*) to assist them in identifying relevant documents.

Subjects were asked to perform the retrieval task as a two-step procedure. They were first provided with the list of LDA topics represented by a given modality (term list, textual label or image), and a query. They were then asked to identify all topics that were potentially to the query. Figure 1 shows the topic browser interface for the three different modalities. In the second step, the participant was presented with a list of documents associated with the selected topics. Documents were presented in random order. Each document was represented by its title, and users were able to read its content in a pop-up window. Figure 2 shows a subset of the documents that are associated with the topics selected in the first step. The documents that are presented to the user in the second step have high conditional probabilities of being associated with the topics that were selected in the first stage. However, it should be noted that this does not guarantee that they are also relevant to any given query.

We also asked users to complete a post-task questionnaire once they had completed the retrieval task. The questionnaire consisted of five questions, which were intended to provide insights into participant satisfaction with the retrieval task and the topic browsing system. Participants assigned an integer score from 1 to 7 in response to each question. First, we asked about the usefulness of the different topic representations, i.e. term list, textual labels and image labels. We also asked about the difficulty level of the task (Ease of Search) and the familiarity of the participants with the queries. The questions were as follows:

- How useful were the term lists in representing topics? (“Usefulness (Term list)”)
- How useful were the textual phrases in representing topics? (“Usefulness (Textual label)”)
- How useful were the images in representing topics? (“Usefulness (Image)”)
- How easy was the task? (“Ease of Search”)
- Did you find the queries easy to understand? (“Query Familiarity”)

3.5 Subjects and Procedure

We recruited 15 members of research staff and graduate students at the University of Sheffield, University of Melbourne and King’s College for the user study. All of the participants had a computer science background, and were also all familiar with on-line digital library and retrieval systems.

Each participant was first asked to sign up to our on-line system, in order to track a given user session

[9] http://koen.me/research/colordescriptors

across time. After logging in, participants had access to a personalised main page where they could read the instructions for the task, see how many queries they have completed so far, or select to perform a new query.

Participants were asked to perform the task for each of the 20 queries, which were presented in random order. The topic representation for each query was randomly chosen, and participants annotated different topics using varying topic representations. Topics and documents were presented in random order to ensure there was no learning effect where participants became familiar with the order and were able to annotate some queries more quickly. We also encouraged participants to perform their allocated queries in multiple sessions by allowing them to return to the interface to complete further queries, provided they completed the overall task within a week.

4 Results

We begin by exploring the number of documents retrieved (Section 4.1) and proportion of retrieved documents that were relevant (Section 4.2). Further analysis is carried out to determine relevance of the retrieved documents based on the topics that were selected in the first stage (Section 4.3). Finally, results from the post-task questionnaire are discussed (Section 4.4).

4.1 Number of Retrieved Documents

We assume that the number of retrieved documents for the three topic browsing systems is indicative of the time required to interpret topics and identify relevant ones. Topic representations that are difficult to interpret will require more time for participants to understand, which will have a direct effect on the number of documents retrieved.

Table 3 shows the number of documents retrieved for each query and modality. Representing topics using lists of terms results in the lowest number of documents retrieved both overall ($1,086$) and for the majority of the queries. The highest number of documents retrieved ($1,264$) occurs when the topics are represented using textual phrase labels. This suggests that textual phrase labels are easier to interpret than the other two representations, thereby allowing participants to identify relevant topics more quickly. The number of documents retrieved for the image representation is slightly higher than the term lists but lower than textual phrase labels.

Query	Term list	Text	Image
Travel & Tourism	22	**33**	17
Domestic Politics (USA)	50	65	**78**
War — Civil War	**61**	31	40
Biographies, Personalities, People	27	**37**	29
Defence	26	**51**	29
Crime, Law Enforcement	34	**49**	25
Religion	84	**97**	44
Disasters & Accidents	**73**	62	63
International Relations	58	**85**	37
Science & Technology	**60**	38	56
Employment/Labour	51	49	**58**
Government Finance	42	**61**	34
Weather	95	**129**	111
Elections	47	**58**	50
Environment & Natural World	33	**69**	41
Arts, Culture, Entertainment	45	**70**	30
Health	**82**	76	37
European Commission (EC) Institutions	48	42	**52**
Sports	113	114	**228**
Welfare, Social Services	35	48	**56**
Total	1,086	$\mathbf{1,264}$	1,115

Table 3: Number of retrieved documents for each query and topic representation.

The number of retrieved documents is high for queries that are associated with many relevant docu-

Query	Term list	Text	Image
Travel & Tourism	**0.73**	0.42	0.59
Domestic Politics (USA)	0.62	**0.69**	**0.69**
War — Civil War	0.82	0.71	**0.90**
Biographies, Personalities, People	0.11	0.14	**0.24**
Defence	0.23	**0.27**	0.07
Crime, Law Enforcement	**0.38**	0.35	0.20
Religion	0.73	0.82	**0.98**
Disasters & Accidents	0.60	0.53	**0.70**
International Relations	0.66	0.69	**0.70**
Science & Technology	0.67	**0.79**	0.73
Employment/Labour	**0.80**	0.76	0.72
Government Finance	0.71	**0.80**	0.53
Weather	**0.79**	0.62	0.62
Elections	0.77	0.48	**0.84**
Environment & Natural World	0.45	**0.54**	0.49
Arts, Culture, Entertainment	0.44	0.04	**0.50**
Health	**0.84**	0.58	0.41
European Commission (EC) Institutions	**0.35**	0.33	0.33
Sports	**0.99**	0.98	0.98
Welfare, Social Services	**0.17**	0.00	0.04
Average	**0.59**	0.53	0.56

Table 4: Precision for each query and topic representation.

ments (*Sports* in term lists, textual phrase labels and image labels; *Domestic Politics (USA)* in image labels). The relatively large number of relevant documents leads to LDA generating a large number of topics relevant to them which, in turn, provides users with many topics through which relevant documents can be selected. In addition, queries such as *Weather* and *Religion* are highly distinct from other queries, making it easier to identify documents relevant to them. On the other hand, the queries for which the fewest documents are retrieved are those that are associated with a small number of relevant documents, i.e. *Travel & Tourism* and *Biographies*.

Further analysis compared the documents retrieved for individual queries. We computed the Pearson's correlation coefficient between the number of documents retrieved for each query across the three topic representations. We observe a high correlation between term lists and textual phrase labels ($r = 0.76$), and term lists and image labels ($r = 0.74$), while the correlation between textual phrase and image labels is lower ($r = 0.63$). These results demonstrate that the topic representation does not strongly affect the relative number of documents retrieved for each query. For example, for all three topic representations, two queries (*Sports* and *Weather*) appear within the top five of the ranking of documents retrieved, and three queries (*Biographies, Personalities, People*; *Crime, Law Enforcement* and *Defence*) appear within the bottom five. Correlation between term lists and textual phrase labels, and term lists and image labels is higher than the correlation between textual phrase and image labels. The main reason might be that both textual phrase and image labels are automatically generated from the topics, which introduces noise.[10] Comparing two noisy methods produces a lower correlation than when just one of them is noisy.

4.2 Precision

We also tested the performance of the different topic representations in terms of the proportion of retrieved documents that are relevant to the query, by computing the average precision for each query across all fifteen users. Results are shown in Table 4. Term lists achieve a higher precision (0.59) than either textual phrase (0.53) or image (0.56) labels. This is somewhat expected since labelling is a type of summarisation, and some loss of information is inevitable. Another possible reason is that the textual phrase and image labels are assigned using auto-

[10] Note that the topics themselves are, of course, automatically generated and potentially noisy, but in terms of topic labelling, constitute the ground truth for a given topic.

matic methods (see Sections 3.3.2 and 3.3.3), which leads to occasional bad label assignments to topics.

Queries such as *Sports*, *Health*, *Religion* and *War — Civil War* are in the top-3 precision for the three topic representations. Identifying relevant documents might be easier for these queries since they tend to be distinct from other queries, making the process of identifying relevant documents more straightforward. On the other hand, we observed low precision for queries that have a low number of relevant documents associated with them such as *Welfare, Social Services* and *Biographies, Personalities, People*.

We computed the Pearson's correlation coefficient between the precisions for the queries across topic representations. An interesting finding is the similarly high correlation achieved between term lists and textual phrase labels ($r = 0.83$), and term lists and image labels ($r = 0.84$). Correlation between textual phrase and image labels is lower ($r = 0.79$) suggesting that there is greater disparity between the queries for which the two methods achieve high/low precision. This is also likely to happen because of bad labelling of topics.

4.3 Document Relevance Based on Topic Selection

We further evaluated the various topic representations by measuring the relevance of the retrieved documents based on the topic selection in the first step of the retrieval task process (see Section 3.4). We define the relevant probability sum as the aggregated probabilities of the topics selected by the participants, given the relevant documents retrieved for each query. In the same fashion, the irrelevant probability sum is computed as the aggregated probabilities of the retrieved documents that are not relevant to the given query. Intuitively, this metric associates retrieved documents with the topics selected for a given query and topic representation. The sum of probabilities for relevant and irrelevant documents for a given query is computed as follows:

$$P_{relevant} = \frac{1}{|U|} \sum_{u \in U} \sum_{d \in D^u_{rel}} \sum_{t \in T_u} P(t|d) \quad (1)$$

$$P_{irrelevant} = \frac{1}{|U|} \sum_{u \in U} \sum_{d \in D^u_{irr}} \sum_{t \in T_u} P(t|d) \quad (2)$$

where d is a document, D^u_{rel} is the set of relevant documents retrieved by a user u, D^u_{irr} is the set of irrelevant documents retrieved, T_u is the set of topics selected by u in the first step of the task, $P(t|d)$ is the conditional probability of topic t given the document d according to the topic model, and U is the set of users who performed the query.

Table 5 shows the results of the average probability sum for relevant and irrelevant documents retrieved by users for each query and topic representation. The results show that both labelling methods perform better than the term list representation for retrieving relevant documents. Textual phrase labels perform best, while image labels obtain comparable performance. Apart from the fact that labelling methods allow users to retrieve more documents, they also allow users to select more relevant topics for a given query.

On the other hand, the probability sum for irrelevant topics selected using the labelling algorithms is higher than term lists. Using lists of terms, participants select a lower number of irrelevant topics, which results in lower irrelevant probability sum. The main reason might be the false labels assigned to topics by these algorithms resulting in irrelevant topic selection by users.

We computed the ratio of the probability masses of the relevant and irrelevant documents retrieved for each topic. The highest ratio (2.5) was obtained when the image labels were used. The ratio for the topic terms is similar (2.3) while the ratio for textual phrases is lower (1.8). This suggests that the topic terms and image labels allow users to identify potentially relevant topics more accurately than when textual labels were used. This is supported by the rankings of the different approaches in terms of their overall precision (see Table 4).

4.4 Post-task Questionnaire

The main finding of the post-task questionnaire is that all of the modalities achieve similar scores in terms of usefulness, as detailed in Table 6. Term lists achieve the highest average score (4.33) while textual phrase labels are close behind (4.26), and image labels slightly lower again (4.00). This demonstrates both that there is room for improvement in all modalities (recalling that the scores are out of 7), and that the different topic representations can be complementary in topic browsers, providing users

Query	Relevant			Irrelevant		
	Term list	Text	Image	Term list	Text	Image
Travel & Tourism	0.00	0.04	0.00	0.00	0.03	0.04
Domestic Politics (USA)	0.29	0.03	0.10	0.04	0.09	0.00
War — Civil War	0.03	0.00	0.15	0.07	0.00	0.03
Biographies, Personalities, People	0.00	0.00	0.00	0.04	0.04	0.04
Defence	0.00	0.01	0.00	0.00	0.05	0.00
Crime, Law Enforcement	0.01	0.05	0.00	0.01	0.18	0.00
Religion	0.18	0.03	0.01	0.10	0.00	0.06
Disasters & Accidents	0.35	0.10	0.26	0.04	0.01	0.03
International Relations	0.04	0.11	0.01	0.04	0.02	0.18
Science & Technology	0.04	0.21	0.07	0.07	0.00	0.02
Employment/Labour	0.06	0.17	0.29	0.00	0.00	0.00
Government Finance	0.00	0.43	0.10	0.02	0.16	0.23
Weather	0.38	0.88	0.33	0.10	0.26	0.00
Elections	0.25	0.06	0.14	0.04	0.04	0.03
Environment & Natural World	0.07	0.59	0.05	0.03	0.19	0.04
Arts, Culture, Entertainment	0.01	0.00	0.00	0.03	0.33	0.00
Health	0.00	0.12	0.00	0.01	0.20	0.03
European Commission (EC) Institutions	0.00	0.09	0.00	0.06	0.00	0.00
Sports	0.08	0.25	1.38	0.00	0.01	0.07
Welfare, Social Services	0.03	0.00	0.00	0.11	0.22	0.36
Average	0.09	0.16	0.15	0.04	0.09	0.06

Table 5: Document relevance based on topic selection.

with alternative ways to explore a document collection.

The participants found the retrieval task quite challenging (3.53), although the average score for Query Familiarity was higher (4.40). Combined, these suggest that the majority of users were reasonably comfortable with the queries and that this is not a likely cause of the lower score for ease of search. Rather, we consider it be reflect the nature of the task and the limited time available for each query.

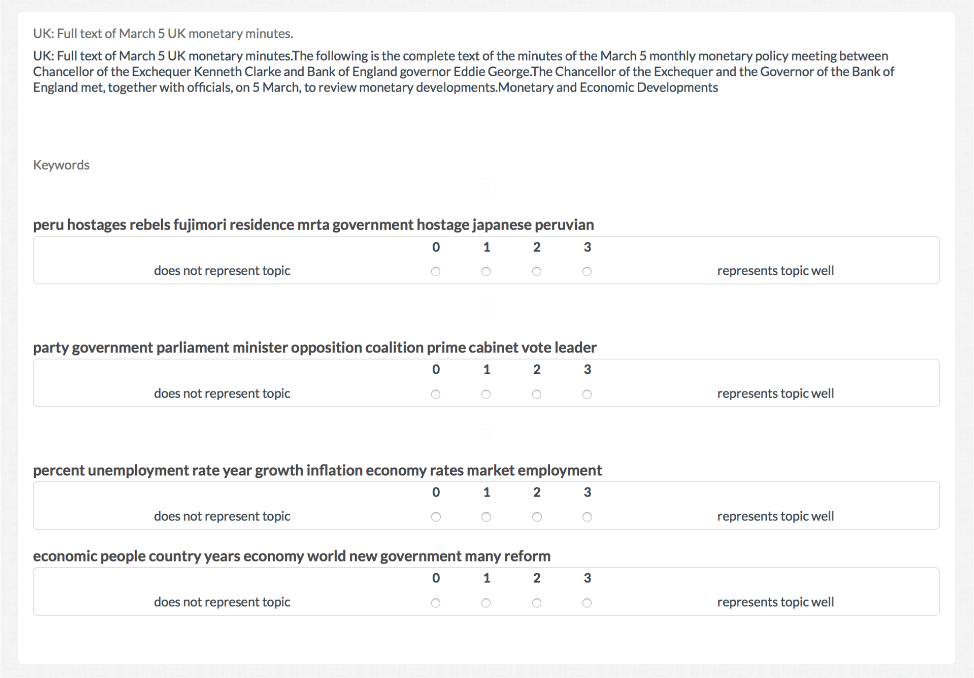
Question	Average
Usefulness (Term list)	4.33
Usefulness (Text)	4.26
Usefulness (Image)	4.00
Query Familiarity	4.40
Easy of Search	3.53

Table 6: Results of the post-task questionnaire.

5 Document Topic Label Relevance

5.1 Human Judgements of Label Relevance

We carried out further analysis to explore the accuracy of the topic labelling methods. A crowdsourc-



UK: Full text of March 5 UK monetary minutes.

UK: Full text of March 5 UK monetary minutes.The following is the complete text of the minutes of the March 5 monthly monetary policy meeting between Chancellor of the Exchequer Kenneth Clarke and Bank of England governor Eddie George.The Chancellor of the Exchequer and the Governor of the Bank of England met, together with officials, on 5 March, to review monetary developments.Monetary and Economic Developments

Keywords

peru hostages rebels fujimori residence mrta government hostage japanese peruvian

	0	1	2	3	
does not represent topic	○	○	○	○	represents topic well

party government parliament minister opposition coalition prime cabinet vote leader

	0	1	2	3	
does not represent topic	○	○	○	○	represents topic well

percent unemployment rate year growth inflation economy rates market employment

	0	1	2	3	
does not represent topic	○	○	○	○	represents topic well

economic people country years economy world new government many reform

	0	1	2	3	
does not represent topic	○	○	○	○	represents topic well

Figure 3: Document topic relevance judgement interface.

ing experiment was carried out in which participants were asked to rate topic labels using an annotation task that is similar to the "intruder detection" task (Chang et al., 2009) used to quantify topic interpretability.

Human judgements of the suitability of each label were obtained using the Crowdflower crowdsourcing platform.[11] The document with the highest marginal probability is identified for each of the 84 topics used in the previous experiment. This document is shown to the annotator together with four labels, one representing the topic and the other three representing randomly-selected topics with low marginal probability for the document. The same three random topics are shown to all annotators for each document (although note that different random topics are used across questions). The order in which the topics are shown to annotators is randomised. Annotators were asked to judge the appropriateness of each topic label from 0 (irrelevant) to 3 (very relevant) with respect to the document's main thematic content. The four topics were represented using each of the three topic modalities, i.e. term lists, text phrases and images, and each topic rated by at least 10 annotators. Figure 3 shows the interface of the crowdsourcing experiment.

This allows us to directly evaluate the interpretability of the topic representations, since we assume that if the topic labels are appropriate then annotators will assign higher scores to labels which are relevant to a document than those which are randomly chosen.

Quality control in crowdsourcing experiments ensures reliability (Kazai, 2011). To avoid random answers, control questions with obvious answers were included in the survey. For example, we presented annotators with a document about finance where the four available labels were a topic about finance and three stop words. Annotations by participants who failed to answer these questions correctly or gave the same rating to all topics were ignored.

5.2 Responses

A total of $2,520$ filtered responses was obtained from 66 participants. The average response for each document–topic pair was calculated in order to create the final similarity judgement. The variance across judges (excluding control questions) was in the range 0.22–0.29.

To measure Inter-Annotator agreement (IAA), we

[11] http://crowdflower.com

first calculated the Spearman's ρ between the ratings given by an annotator and the average ratings from all other annotators for those same document–topic pairs. We then averaged the ρ across annotators and document–topic pairs. Average IAA scores are shown in Table 7. The lower agreement for the image labels indicates that the annotators found it more difficult to identify the correct label.

5.3 Evaluation

The topic representations were analysed using the following two metrics:

- **Top-1 average rating:** the average human rating assigned to each topic label. This provides an indication of the overall quality of the labels the annotators judge as the best one. The highest possible score averaged across all topics is 3.

- **Match@1:** the relative frequency of the correct topic for a given representation being rated the highest out of the four topics.

Results are shown in Table 8. Term lists achieve the best performance for both the Top-1 Average and Match@1 measures, with scores of 1.70 and 0.92 respectively. As discussed above, term lists have the advantage of being more descriptive and informative since they consist of more words than textual phrase labels. The average ratings assigned by annotators are lower than the average scores assigned by humans to textual phrase and image labels in similar crowdsourcing experiments (Lau et al., 2011; Aletras and Stevenson, 2013b). This is due to our labelling task being different in nature. We asked annotators to judge the appropriateness of the label given a document with high probability for that topic while previous experiments (Lau et al., 2011; Aletras and Stevenson, 2013b) seek to find the appropriateness of the label given the term list for a topic.

Textual phrase labels also perform well, with annotators able to identify the correct topic 83% of the time. Scores for this representation are close to those for the term lists despite the verbosity of topic labels generally being much lower than term lists. The average length of the textual phrase labels used in the experiment was 2.7 words while term lists contained 10 words. It is possible that the performance of textual phrase labels may equal, or even exceed, that of term lists with better labelling algorithms.

Representation	IAA
Term list	0.81
Text	0.78
Image	0.57

Table 7: IAA across the four topic labels and document–topic pairs.

Representation	Top-1 Average	Match@1
Term list	1.70	0.92
Text	1.57	0.83
Image	0.83	0.67
Upper Bound	3.0	1.0

Table 8: Results for the document topic detection task.

On the other hand, results for image labels are substantially lower (Top-1 Average = 0.83, and Match@1 = 0.67). This suggests that the image labels are not as clear as the other two types, making it difficult for annotators to identify the correct one. Image labels are also generated automatically and mistakes in this process are likely to explain the lower performance to some extent. However, it is also possible that images are inherently more ambiguous than the other two types of labels, making it difficult for annotators to identify the correct topic.

The results from this experiment indicate some variation between how effectively the three topic representations are able to convey the semantics of a topic. However, results from the exploratory search experiment (Section 4) suggest that any of the three are useful ways of representing documents within a collection and, in particular, allow relevant documents to be identified. Term lists provide a faithful representation of a topic, since they are generated directly from its keywords, while the textual phrase and image labels are generated using labelling algorithms which rely on external resources and may make errors. On the other hand, the textual phrase and image labels are more compact than term lists, allowing them to be interpreted more quickly and more to be fitted onto an interface. It is likely that these factors (fidelity and verbosity) balance out when the topic representations are used in the exploratory search interface. It is also possible, of course, that performance using textual phrase or image labels could be improved with the development of more accurate labelling algorithms.

6 Conclusion

We compared three representations for automatically-generated topics: (1) lists of terms, (2) textual phrase labels, and (3) image labels. These representations were compared within an exploratory browsing interface and an experiment was carried out in which users were asked to retrieve relevant documents using the interface.

Results show that participants were able to identify relevant documents using any of the three topic representations. They were able to identify more documents when labels were used to represent topics than when term lists were used, suggesting that participants can interpret labels more quickly. However, a greater proportion of the retrieved documents are relevant to the query for term lists than either type of label, suggesting that term lists contain more accurate information than the labels. This hypothesis was explored in a further experiment in which participants were asked to identify the most appropriate topics for documents. The information in term lists was found to be more accurate, which is to be expected since the labels are effectively summaries of the topics and, since they are generated automatically from the topics, inevitably contain some errors (Lau et al., 2011; Aletras and Stevenson, 2013b). Despite this, the number of relevant documents retrieved in the exploratory search experiment is very similar for all approaches. Overall, textual phrases and image labels can be interpreted more quickly than term lists but not as accurately.

Results indicate that automatically generated labels are a suitable way for representing topics within search interfaces. They have the advantage of being more compact than the term lists that are normally used, providing greater flexibility in the creation of exploratory interfaces. Retrieval performance is comparable to when term lists are used and is likely to increase with improved topic labelling methods.

In the future, we would like to make use of other digital library collections to find out how successful these techniques are in other domains. We would also like to explore the connection between improved labelling methods and task performance.

Acknowledgements

The authors are grateful for support received from the Engineering and Physical Sciences Research Council via the Network on Vision and Language and from the Defence Science & Technology Laboratory through the Centre for Defence Enterprise.

References

Nikolaos Aletras and Mark Stevenson. 2013a. Evaluating topic coherence using distributional semantics. In *Proc. of the 10th Int. Conf. on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany.

Nikolaos Aletras and Mark Stevenson. 2013b. Representing topics using images. In *Proc. of the 2013 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies*, pages 158–167, Atlanta, Georgia, USA.

Nikolaos Aletras and Mark Stevenson. 2014. Labelling Topics using Unsupervised Graph-based Methods. In *Proc. of the 52nd Annual Meeting of the Assoc. for Computational Linguistics (Volume 2: Short Papers)*, pages 631–636, Baltimore, Maryland.

David M. Blei and Michael I. Jordan. 2003. Modeling annotated data. In *Proc. of the 26th Annu. Int. ACM SIGIR Conf. on Research and Development in Inform. Retrieval (SIGIR 03)*, pages 127–134, Toronto, Canada.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. of Mach. Learning Research*, 3:993–1022.

Cano Basave, A. Elizabeth, He, Yulan and Xu, Ruifeng. 2014. Automatic Labelling of Topic Models Learned from Twitter by Summarisation. In *Proc. of the 52nd Annual Meeting of the Assoc. for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, Baltimore, Maryland.

Allison June-Barlow Chaney and David M. Blei. 2012. Visualizing topic models. In *Proc. of the 6th Int. AAAI Conf. on Weblogs and Social Media*, pages 419–422, Dublin, Ireland.

Jonathan Chang, Jordan Boyd-Graber, and Sean Gerrish. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Inform.*, pages 1–9.

Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: designing model-driven visualizations for text analysis. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Syst.*, pages 443–452. ACM.

Christopher Collins, Fernanda B Viegas, and Martin Wattenberg. 2009. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. of IEEE Sympos. on Visual Analytics Sci. and Technology (VAST 2009)*, pages 91–98. IEEE.

Debasis Ganguly, Manisha Ganguly, Johannes Leveling, and Gareth J.F. Jones. 2013. TopicVis: A GUI for Topic-based feedback and navigation. In *Proc. of the 36th Annu. Int. ACM SIGIR Conf. on Research and Develop. in Inform. Retrieval (SIGIR 13)*, pages 1103–1104, Dublin, Ireland.

Matthew J Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. The Topic Browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*, Whistler, Canada.

Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. 2012. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Trans. on Intelligent Syst. Technology*, 3(2):23:1–23:26.

Karl Grieser, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg. 2011. Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness. *J. on Computing and Cultural Heritage (JOCCH)*, 3(3):10:1–10:20.

Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. 2003. An analytical comparison of approaches to personalizing PageRank. Tech. Rep. 2003–35, Stanford InfoLab.

Marti A. Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. of the ACM*, 49(4):59–61.

Marti A. Hearst. 2009. *Search User Interfaces*. Cambridge University Press, Cambridge, UK.

Alexander Hinneburg, Rico Preiss, and René Schröder. 2012. TopicExplorer: Exploring document collections with topic models. In Peter A. Flach, Tijl Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases, volume 7524 of Lecture Notes in Comput. Sci.*, pages 838–841. Springer, Heidelberg, Germany.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proc. of the 22nd Annu. Int. ACM SIGIR Conf. on Research and Develop. in Inform. Retrieval (SIGIR 99)*, pages 50–57, Berkeley, California, United States.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using DBpedia. In *Proc. of the 6th ACM Int. Conf. on Web Search and Data Mining (WSDM 13)*, pages 465–474, Rome, Italy.

Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. *Advances in Information Retrieval*, pages 165–176.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for topic labelling. In *Proc. of the 23rd Int. Conf. on Computational Linguistics (COLING 10)*, pages 605–613, Beijing, China.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proc. of the 49th Annu. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA.

David G. Lowe. 1999. Object Recognition from Local Scale-invariant Features. In *Proceedings of the 7th IEEE Int. Conf. on Comput. Vision*, pages 1150–1157, Kerkyra, Greece.

David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. of Comput. Vision*, 60(2):91–110.

Davide Magatti, Silvia Calegari, Davide Ciucci, and Fabio Stella. 2009. Automatic Labeling of Topics. In *Proc. of the 9th Int. Conf. on Intelligent Systems Design and Applications (ICSDA 09)*, pages 1227–1232, Pisa, Italy.

Xian-Li Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. 2012. Automatic labeling hierarchical topics. In *Proc. of the 21st ACM Int. Conf. on Inform. and Knowledge Manage. (CIKM 12)*, Maui, Hawai, USA.

Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. of the ACM*, 49(4):41–46.

Qiaozhu Mei and ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proc. of the 11th ACM Int. Conf. on Knowledge Discovery in Data Mining (SIGKDD 05)*, pages 198–207, Chicago, Illinois, USA.

Qiaozhu Mei, Xuehua Shen, and Cheng Xiang Zhai. 2007. Automatic Labeling of Multinomial Topic Models. In *Proc. of the 13th ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD 07)*, pages 490–499, San Jose, California, USA.

David Newman, Timothy Baldwin, Lawrence Cavedon, Eric Huang, Sarvnaz Karimi, David Martinez, Falk Scholer, and Justin Zobel. 2010. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):169–175.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Tech. Rep. 1999–66, Stanford InfoLab.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus volume 1 from yesterdays news to tomorrows language resources. In *Proc. of the 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002)*, pages 827–832, Las Palmas, Canary Islands.

Greg Smith, Mary Czerwinski, B Robbins Meyers, G Robertson, and DS Tan. 2006. Facetmap: A scalable search and browse visualization. *IEEE Trans. Vis. Comput. Graphics*, 12(5):797–804.

Justin Snyder, Rebecca Knowles, Mark Dredze, Matthew Gormley, and Travis Wolfe. 2013. Topic models and metadata for visualizing text corpora. In *Proc. of the 2013 North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies–Demonstration Session*, pages 5–9, Atlanta, Georgia.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proc of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 12)*, pages 952–961, Jeju Island, Korea.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *J. of the American Statistical Assoc.*, 101(476):1566–1581.

Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: a visual exploratory text analytic system. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 153–162, Washington DC, USA.