Evaluating Topic Representations for Exploring Document Collections

Nikolaos Aletras
Computer Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom. E-mail: nikos.aletras@gmail.com

Timothy Baldwin
Computing and Information Systems, University of Melbourne, CIS, Melbourne, Victoria 3204, Australia. E-mail: tb@ldwin.net

Jey Han Lau
Department of Philosophy, King's College, Strand, London, London, WC2R 2LS, United Kingdom. E-mail: jeyhan.lau@gmail.com

Mark Stevenson
Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S Yorks S1 4DP, United Kingdom. E-mail: mark.stevenson@sheffield.ac.uk

Topic models have been shown to be a useful way of representing the content of large document collections, for example, via visualization 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; that is the top-*n* words with highest conditional probability within the topic. Other topic representations such as textual and image labels also have been proposed. However, there has been no comparison of these alternative representations. In this article, we compare 3 different topic representations in a document retrieval task. Participants were asked to retrieve relevant documents based on predefined queries within a fixed time limit, presenting topics in one of the following modalities: (a) lists of terms, (b) textual phrase labels, and (c) image labels. Results show that textual labels are easier for users to interpret than are 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 labeling methods are an effective alternative topic representation.

Introduction

In recent years, a large amount of information has been made available online in digital libraries, collections, and archives. Much of this information is stored in unstructured format (e.g., text) and is not organized 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, Viegas, & Wattenberg, 2009; Hearst, 2006; Smith, Czerwinski, Meyers, Robertson, & Tan, 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 & Jordan, 2003; Blei, Ng, & Jordan, 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 visualization interfaces (topic browsers) (Chaney & Blei, 2012; Ganguly, Ganguly, Leveling, & Jones, 2013; Gretarsson et al., 2012; Hinneburg, Preiss, & Schröder, 2012; Snyder, Knowles, Dredze, Gormley, & Wolfe, 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; that is, 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 automatically generate these (Lau, Newman, Karimi, & Baldwin, 2010; Lau, Grieser, Newman, & Baldwin, 2011; Mei, Shen, & Zhai, 2007). Approaches that make use of alternative modalities, such as images (Aletras & Stevenson, 2013b), also have 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 does the standard term list approach. However, there has not, to our knowledge, been any empirical validation of this intuition—a shortcoming that this article aims to address—in carrying out a task-based evaluation of different topic model representations. In this, we compare three approaches to representing topics: (a) a standard term list, (b) textual phrase labeling, and (c) image labeling. These are used to represent topics generated from a digital archive of newswire 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:

RQ1. Which topic representations are suitable within a document browser interface?
RQ2. What is the impact of different topic representations on human search effectiveness for a given query?

First, we review previous work on automatically labeling topics and the use of topic models to create search interfaces. Then, we introduce an experiment in which three approaches to topic labeling are applied and evaluated within an exploratory search interface. The results of the experiment on exploratory search are presented next, followed by intrinsic evaluation of the labels generated by the different methods.

Related Work

In early research on topic modeling, 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 & Zhai, 2005; Teh, Jordan, Beal, & Blei, 2006).

The first attempt to automatically assigning labels to topics was 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 minimizes 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, Calegari, Ciucci, and Stella (2009) introduced an approach for labeling topics that relies on two manually labeled hierarchical knowledge resources: the Google Directory and the OpenOffice English Thesaurus. The automatic labelling of topics algorithm computes the similarity between latent Dirichlet allocation (LDA)-inferred topics and categories in the topic tree, a preexisting hierarchical set of labeled 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 reranking model.

Lau et al. (2011) proposed a method for automatically labeling 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, Baldwin, Bohnert, & Sonenberg, 2011) and removed. Finally, the top-five 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 labeling 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, Hayes, Karnstedt, and Greene (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.

Basave et al. (2014) presented a method for labeling LDA topics trained on social media streams (i.e., Twitter) using summarization 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, Brin, Motwani, & Winograd, 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 labeling 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, and visual features extracted from the analysis of the images themselves.

Topic modeling has been used to support browsing in large document collections (Chaney & Blei, 2012; Chuang, Ramage, Manning, & Heer, 2012; Ganguly et al., 2013; Gardner et al., 2010; Hinneburg et al., 2012; Newman et al., 2010; Snyder et al., 2013; Wei et al., 2010). 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 previous research on automatically generating labels for topics. We address this limitation by making use of three approaches to labeling topics within a topic-based browser and carrying out experiments to compare their effectiveness.

---

[1]http://dbpedia.org

Methods

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

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

Document Collection

We make use of a subset of the Reuters Corpus (Rose, Stevenson, & Whitehead, 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 judgments to determine the accuracy of the documents retrieved by users.

Twenty topic categories were selected, and 100,000 documents were randomly extracted from the Reuters Corpus. Each document is preprocessed by tokenization, 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.

TABLE 1. Number of documents in each Reuters Corpus topic category.

Reuters Topic Category (Query)	No. of Documents
Travel & Tourism	314
Domestic Politics (USA)	27,236
War—Civil War	16,615
Biographies, Personalities, People	2,601
Defense	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 2. Labels generated for an example topic. [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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

Topic Modeling

An LDA model was trained[2] over the document collection using variational inference (Blei & Jordan, 2003). The number of topics learned was set to $T = 100$ since topic interpretability in LDA tends to stabilize when $T \geq 100$ (Stevens, Kegelmeyer, Andrzejewski, & Buttler, 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.

Topic Browsing Systems

The topic-browsing system developed for this study is based on the publicly available Topic Model Visualization Engine (TMVE; Chaney & Blei, 2012). TMVE uses a document collection and an LDA model trained over that collection (discussed earlier). It generates a topic-browsing system with three main components: (a) a main page, (b) topic pages, and (c) 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: (a) term lists, (b) textual phrase labels, and (c) images. The term lists are created using a standard approach (discussed later), the textual phrase labels are generated from Wikipedia article titles (Lau et al., 2011) (discussed later), and the image labels are generated using publicly available images from Wikipedia (Aletras & Stevenson, 2013b) (discussed later). By default, TMVE only 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.

Term lists. Term lists are generated using the default approach of TMVE: 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 modeling research community.

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-seven topic terms[4] to search Wikipedia using Wikipedia's native search application program interface (API) and Google's site-restricted search. We collect the top-eight article titles returned from each of the search engines;[5] these constitute the primary 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 Related Article Conceptual Overlap (RACO) lexical association method (Grieser et al., 2011). The component n-grams that pass the RACO filter constitute the secondary candidates. Last, we include the top-five 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 the 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.

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 license. The top-five terms from a topic are used to query Bing using

---

its Search API.[7] The search is restricted to the 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 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 tokenization and removal of stop words.

Visual information is extracted using low-level image keypoint descriptors; that is, scale-invariant feature transform (SIFT) features (Lowe, 1999, 2004) sensitive to color information. Image features are extracted using dense sampling and described using Opponent color 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, Kamvar, & Jeh, 2003) is used to rank the candidate images. The personalization vector of PPR is initialized 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.

Task

The aim of the task was to identify as many documents as possible that are relevant to a set of queries. 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 also were 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.

Participants were asked to perform the retrieval task as a two-step procedure. They first were provided with the list of LDA topics represented by a given modality (term list, textual label, or image), and a query. Next, they were asked to identify all topics that were potentially related to the query. Figure 1 shows the topic browser interface for the three different modalities. In the second step, participants were 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, note that this does not guarantee that they also are relevant to any given query.

In addition, we asked users to complete a posttask 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 (ranging from *useful/ easy/familiar* to *very useful/easy/familiar*) 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")

Participants and Procedure

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

Each participant was first asked to sign up to our online system so we could track a given user session across time. After logging in, participants had access to a personalized main page where they could read the instructions for the task, see how many queries they had 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. 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 that there was no learning effect where participants became familiar with the order and were able to more quickly annotate some queries. 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 that they completed the overall task within 1 week.

Results

We begin by exploring the number of documents retrieved and the proportion of retrieved documents that

FIG. 1.   Topic browsing interfaces. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

were relevant. Further analysis is carried out to determine relevance of the retrieved documents based on the topics that were selected in the first stage. Finally, results from the posttask questionnaire are discussed.

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 are the other two representations, thereby allowing participants to more quickly identify relevant topics. The number of documents retrieved for the image representation is slightly higher than the term lists, but lower than textual phrase labels.

The number of retrieved documents is high for queries that are associated with many relevant documents (*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 (*Travel & Tourism* and *Biographies*).

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 ☐

FIG. 1.   *Continued*

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 *Defense*) appear within the bottom five. The 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.

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 15 users. The results are shown in Table 4. Term lists achieve a higher precision (0.59) than did either the textual phrase (0.53) or image (0.56) labels. This is somewhat expected since labeling is a type of summarization, and some loss of information is inevitable. Another possible reason is that the textual

---

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

FIG. 1. *Continued*

phrase and image labels are assigned using automatic methods (discussed earlier), which leads to occasional bad label assignments to topics.

Queries such as *Sports*, *Health*, *Religion,* and *War—Civil War* are in the top-three 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 also is likely to happen because of bad labeling of topics.

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 (discussed earlier). 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:
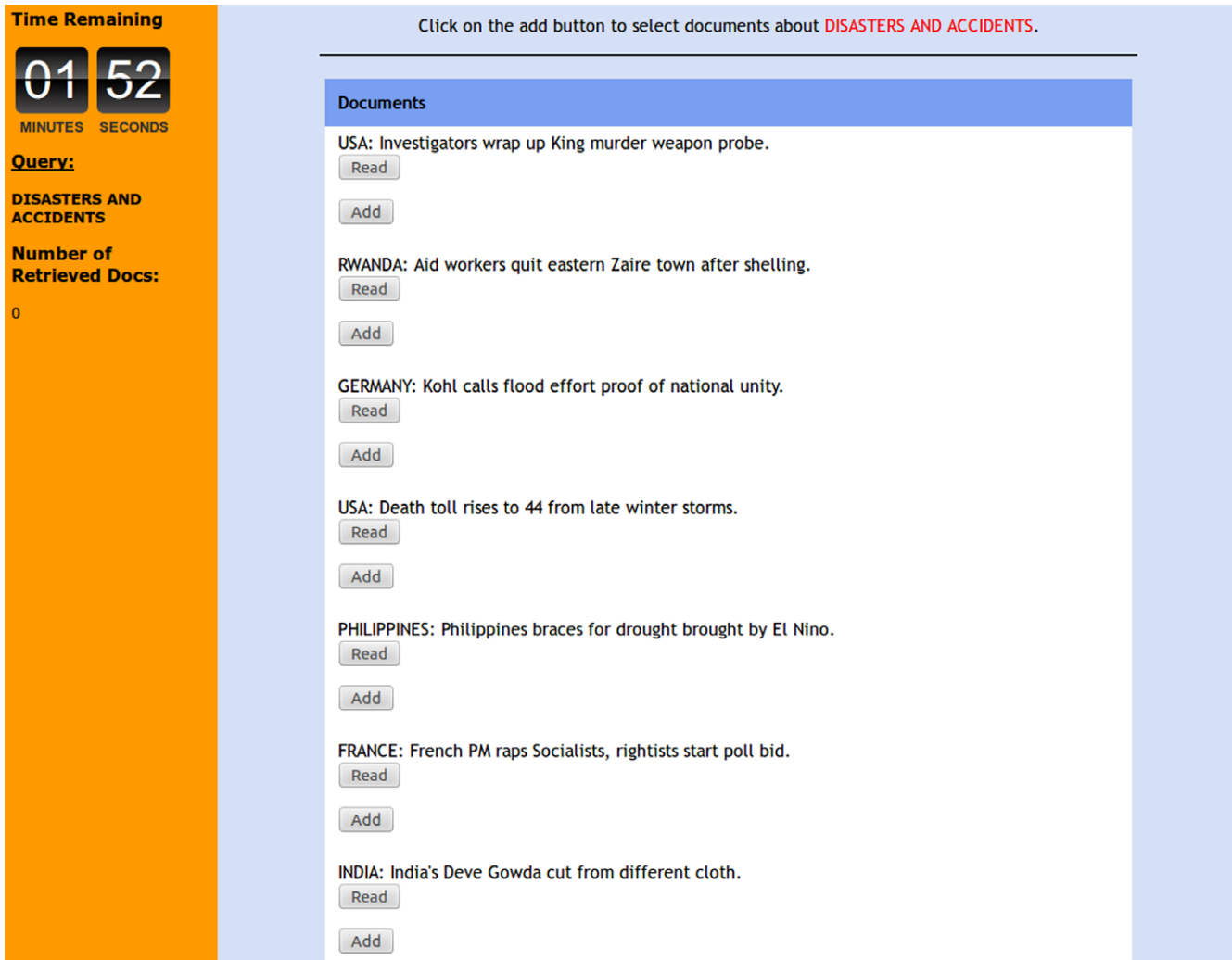
FIG. 2. Topic browsing: List of documents. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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

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

where d is a document, D_{rel}^u is the set of relevant documents retrieved by a user u, D_{irr}^u is the set of irrelevant documents retrieved, T_u is the set of topics selected by user 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 labeling 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 labeling 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 labeling algorithms is higher than are the term lists. Using lists of terms, participants select a lower number of irrelevant topics, which results in a 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).

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

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	**1,264**	1,115

TABLE 4. Precision for each query and topic representation.

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**
Defense	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

Posttask Questionnaire

The main finding of the posttask 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 that there is room for improvement in all modalities (recalling that the scores are of 7 in total) and that

the different topic representations can be complementary in topic browsers, providing users with alternative ways to explore a document collection.

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 the cause of the lower score for ease of search. Rather, we consider it to be reflective of the nature of the task and the limited time available for each query.

Document Topic Label Relevance

Human Judgments of Label Relevance

We conducted further analysis to explore the accuracy of the topic labeling methods. A crowdsourcing 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 judgments 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 randomized. 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 was 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 to 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 in which the four available labels were a topic about finance and three stop words. Annotations by participants who failed to correctly answer these questions or gave the same rating to all topics were ignored.

---

[11]http://crowdflower.com

TABLE 5. Document relevance based on topic selection.

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 6. Results of the posttask questionnaire.

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

Responses

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

To measure inter-annotator agreement (IAA), we first calculated 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.

Evaluation

Topic representations were analyzed 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 that 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 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 earlier, term lists have the advantage of being more descriptive and informative since they consist of more words than do 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 crowd-sourcing experiments (Aletras & Stevenson, 2013b; Lau et al., 2011). This is due to our labeling 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 (Aletras & Stevenson, 2013b; Lau et al., 2011) 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 labeling algorithms.

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 are the

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

FIG. 3. Document topic relevance judgement interface.

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

Representation	IAA
Term list	0.81
Text	0.78
Image	0.57

TABLE 8. Results for the document topic detection task.

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

other two types, making it difficult for annotators to identify the correct one. Image labels also are generated automatically, and mistakes in this process are likely to explain the lower performance to some extent. However, it also is possible that images are inherently more ambiguous than are 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 (discussed earlier) 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 labeling algorithms which rely on external resources and may make errors. On the other hand, the textual phrase and image labels are more compact than are 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 also is possible, of course, that performance using textual phrase or image labels could be improved with the development of more accurate labeling algorithms.

Conclusion

We compared three representations for automatically generated topics: (a) lists of terms, (b) textual phrase labels,

and (c) 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 for either type of label, suggesting that term lists contain more accurate information than do 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 (Aletras & Stevenson, 2013b; Lau et al., 2011). 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 can 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 are 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 labeling 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 also would like to explore the connection between improved labeling methods and task performance.

Acknowledgments

References

Aletras, N., & Stevenson, M. (2013a). Evaluating topic coherence using distributional semantics. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Long Papers (pp. 13–22). Potsdam, Germany: Association for Computational Linguistics.

Aletras, N., & Stevenson, M. (2013b). Representing topics using images. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 158–167). Atlanta, GA: Association for Computational Linguistics.

Aletras, N., & Stevenson, M. (2014). Labelling topics using unsupervised graph-based methods. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Vol. 2: Short Papers) (pp. 631–636). Baltimore, MD: Association for Computational Linguistics.

Basave, C., Elizabeth, A., He, Y., & Xu, R. (2014). Automatic labelling of topic models learned from twitter by summarisation. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Vol. 2: Short Papers) (pp. 618–624). Baltimore, MD: Association for Computational Linguistics.

Blei, D.M., & Jordan, M.I. (2003). Modeling annotated data. In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Inform. Retrieval (SIGIR 03) (pp. 127–134). Toronto, Canada: ACM.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Chaney, A.J.-B., & Blei, D.M. (2012). Visualizing topic models. In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (pp. 419–422). Dublin, Ireland.

Chang, J., Boyd-Graber, J., & Gerrish, S. (2009). Reading tea leaves: How humans interpret topic models. *Neural Inform.*, pp. 1–9.

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: designing model-driven visualizations for text analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 443–452). Austin, Texas: ACM.

Collins, C., Viegas, F.B., & Wattenberg, M. (2009). Parallel tag clouds to explore and analyze faceted text corpora. In Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST 2009) (pp. 91–98). Atlantic City, New Jersey: IEEE.

Ganguly, D., Ganguly, M., Leveling, J., & Jones, G.J.F. (2013). TopicVis: A GUI for Topic-based feedback and navigation. In Proceedings of the 36th annual International ACM SIGIR Conference on Research and Develop. in Information Retrieval (SIGIR 13) (pp. 1103–1104). Dublin, Ireland: ACM.

Gardner, M.J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., & Seppi, K. (2010). The Topic Browser: An interactive tool for browsing topic models. In NIPS Workshop on Challenges of Data Visualization (pp. 5228–5235). Canada: Whistler.

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. (2012). TopicNets: Visual analysis of large text corpora with topic modeling. ACM Transactions on Intelligent System Technology, 3(2), 23, 26.

Grieser, K., Baldwin, T., Bohnert, F., & Sonenberg, L. (2011). Using ontological and document similarity to estimate museum exhibit relatedness. Journal on Computing and Cultural Heritage (JOCCH), 3(3), 10, 20.

Haveliwala, T., Kamvar, S., & Jeh, G. (2003). An analytical comparison of approaches to personalizing PageRank. Tech. Rep. 2003–35, Stanford InfoLab.

Hearst, M.A. (2006). Clustering versus faceted categories for information exploration. Communications of the ACM, 49(4), 59–61.

Hearst, M.A. (2009). Search User Interfaces. Cambridge, United Kingdom: Cambridge University Press.

Hinneburg, A., Preiss, R., & Schröder, R. (2012). TopicExplorer: Exploring document collections with topic models. In P.A. Flach, T. Bie, & N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases, volume 7524 of Lecture Notes in Computer Science (pp. 838–841). Heidelberg, Germany: Springer.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99) (pp. 50–57). Berkeley, CA: ACM.

Hulpus, I., Hayes, C., Karnstedt, M., & Greene, D. (2013). Unsupervised graph-based topic labelling using DBpedia. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM 13) (pp. 465–474). Rome, Italy: ACM.

Kazai, G. (2011). Search of quality in crowdsourcing for search engine evaluation. *Advances in Information Retrieval*, pp. 165–176.

Lau, J.H., Newman, D., Karimi, S., & Baldwin, T. (2010). Best topic word selection for topic labelling. In Proceedings of the 23rd International

Conference on Computational Linguistics (COLING 10) (pp. 605–613). Beijing, China: Coling 2010 Organizing Committee.

Lau, J.H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labelling of topic models. In Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 1536–1545). Portland, OR: Association for Computational Linguistics.

Lowe, D.G. (1999). Object Recognition from Local Scale-invariant Features. In Proceedings of the 7th IEEE International Conference on Computer Vision (pp. 1150–1157). Kerkyra, Greece: IEEE.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. The International Journal of Computer Vision, 60(2), 91–110.

Magatti, D., Calegari, S., Ciucci, D., & Stella, F. (2009). Automatic labeling of topics. In Proceedings of the 9th International Conference on Intelligent Systems Design and Applications (ICSDA 09) (pp. 1227–1232). Pisa, Italy: IEEE.

Mao, X.-L., Ming, Z.-Y., Zha, Z.-J., Chua, T.-S., Yan, H., & Li, X. (2012). Automatic labeling hierarchical topics. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 12) (pp. 2383–2386). Maui, HI: ACM.

Marchionini, G. (2006). Exploratory search: From finding to understanding. Communications of the ACM, 49(4), 41–46.

Mei, Q., & Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the 11th ACM International Conference on Knowledge Discovery in Data Mining (SIGKDD 05) (pp. 198–207). Chicago, IL: ACM.

Mei, Q., Shen, X., & Zhai, C.X. (2007). Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 07) (pp. 490–499). San Jose, CA: ACM.

Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martinez, D., . . . Zobel, J. (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.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Tech. Rep. 1999–66, Stanford InfoLab.

Rose, T., Stevenson, M., & Whitehead, M. (2002). The Reuters Corpus Volume 1 from yesterday's news to tomorrow's language resources. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002) (pp. 827–832). Las Palmas, Canary Islands.

Smith, G., Czerwinski, M., Meyers, B.R., Robertson, G., & Tan, D.S. (2006). Facetmap: A scalable search and browse visualization. IEEE Transactions on Visualization and Computer Graphics, 12(5), 797–804.

Snyder, J., Knowles, R., Dredze, M., Gormley, M., & Wolfe, T. (2013). Topic models and metadata for visualizing text corpora. In Proceedings of the 2013 North American Chapter of the Association for Computational Linguistics: Human Language Technologies–Demonstration Session (pp. 5–9). Atlanta, GA: Association for Computational Linguistics.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 12) (pp. 952–961). Jeju Island, Korea: Association for Computational Linguistics.

Teh, Y.W., Jordan, M.I., Beal, M.J., & Blei, D.M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476), 1566–1581.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M.X., Qian, W., . . . Zhang, Q. (2010). Tiara: a visual exploratory text analytic system. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 153–162). Washington DC, USA: ACM.