Representing Topics Labels for Exploring Digital Libraries

Nikolaos Aletras
Computer Science
University of Sheffield
n.aletras@sheffield.ac.uk

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 set of keywords, i.e. the top-n words with highest marginal probability within the topic. However, alternative topic representations have been proposed, including textual and image labels. In this paper, we compare different topic representations, i.e. sets of topic words, textual phrases and images, in a document retrieval task. We asked participants to retrieve relevant documents based on pre-defined queries within a fixed time limit, presenting topics in one of the following modalities: (1) sets of keywords, (2) textual labels, and (3) image labels. Our results show that textual labels are easier for users to interpret than keywords and image labels. Moreover, the precision of retrieved documents for textual and image labels is comparable to the precision achieved by representing topics using sets of keywords, demonstrating that labelling methods are an effective alternative topic representation.

Keywords

topic model, information retrieval, evaluation

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 know the appropriate keywords, which 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. These problems can be avoided by automatically using large-scale data-analysis techniques to interpret the information and provide it to the user in an easily understandable format.

Other types of interaction, such as exploratory search [1] and sense-making [2], are also important and more suitable when the user is not familiar with the collection. However, users tend to be conservative and resistant to these more experimental modes of interaction. Approaches such as faceted search have proved useful for exploratory search [3, 4, 5], but these presuppose a consistent classification scheme, which does 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). Manual classification is impractical for all but the smallest of collections. An alternative approach is to carry out an automatic analysis of the collection and use the results to create a structure that can be used for browsing.

Topic models [6, 7] 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) [8, 9, 10, 11, 12]. 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, as unsupervised algorithms, can be applied at low cost.

The standard way to represent a topic is using a set of keywords, 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 labels

978-1-4799-5569-5/14/$31.00 ©2014 Crown.

(e.g. EDUCATION for our example topic), can potentially assist with the interpretations of topics, and researchers have developed methods to generate these automatically [13, 14, 15]. Approaches that make use of alternative modalities, such as images [16], have also been proposed.

Intuitively, labels represent topics in a more accessible manner than the standard keyword 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 keyword list, (2) textual labelling, and (3) image labelling. These are used to represent topics generated from a digital library containing archive 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 and conclusions are presented in Sections 4 and 5.

2. RELATED WORK

In early research on topic modelling, topics were represented as lists of keywords with the highest probability, and textual labels were sometimes manually assigned to topics for convenience of presentation of research results [17, 18].

The first attempt to automatically assigning labels to topics is described by Mei et al. [13]. 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. [19] introduced an approach for labelling topics that relies on two hierarchical knowledge resources labelled by humans: the Google Directory and the OpenOffice English Thesaurus. A *topic tree* is a pre-existing hierarchical set of labelled topics. The Automatic Labelling Of Topics algorithm computes the similarity between LDA-inferred topics and topics in a *topic tree* by computing scores using six standard similarity measures. The label for the most similar topic in the *topic tree* is assigned to the LDA topic.

Lau et al. [14] proposed selecting the most representative word 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.

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

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

More recently, Lau et al. [15] 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 n-grams that exist as Wikipedia articles. Outlier labels are identified using a word similarity measure [20], 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.

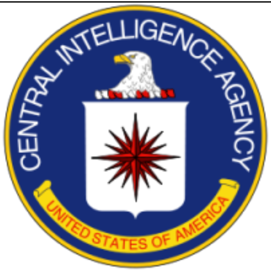
Mao et al. [21] introduced a method for labelling hierarchical topics which makes use of sibling and parent–child relations of topics. Candidate labels are generated using a similar approach to the one used by Mei et al. [13]. Each candidate label is then assigned a score by creating a distribution based 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. [22] 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.

In contrast, Aletras and Stevenson [16] proposed a method for labelling topics using images rather than text. A set of

[1] http://dbpedia.org

Table 2: Labels generated for an example topic.

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

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 [23]. 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 [24, 25, 26, 8, 11, 27, 9, 12]. 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 sets of keywords 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.

3. METHODOLOGY

We conducted a retrieval task to compare three topic representations: (1) lists of keywords, (2) textual labels, and (3) image labels.

3.1 Document Collection

We make use of a subset of the Reuters Corpus [28], 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

We make use of the implementation provided by David Blei[2] to train an LDA model over the document collection using variational inference [29]. The number of topics learned is set to $T = 100$; default settings are used elsewhere.

We choose to generate this number of topics since topic interpretability in LDA becomes stable when $T \geq 100$ [30]. Finally, we removed topics that are difficult to interpret [31] to leave 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) [8]. The TMVE uses a document collection and an LDA model trained over that collection (see above). It generates a topic browsing system with three main components: a main page, topic pages and document pages. The main page contains the list of topics generated. Each topic page shows a list of documents with the highest marginal probability given that topic. Document pages show the content of a document together with its topic distribution.

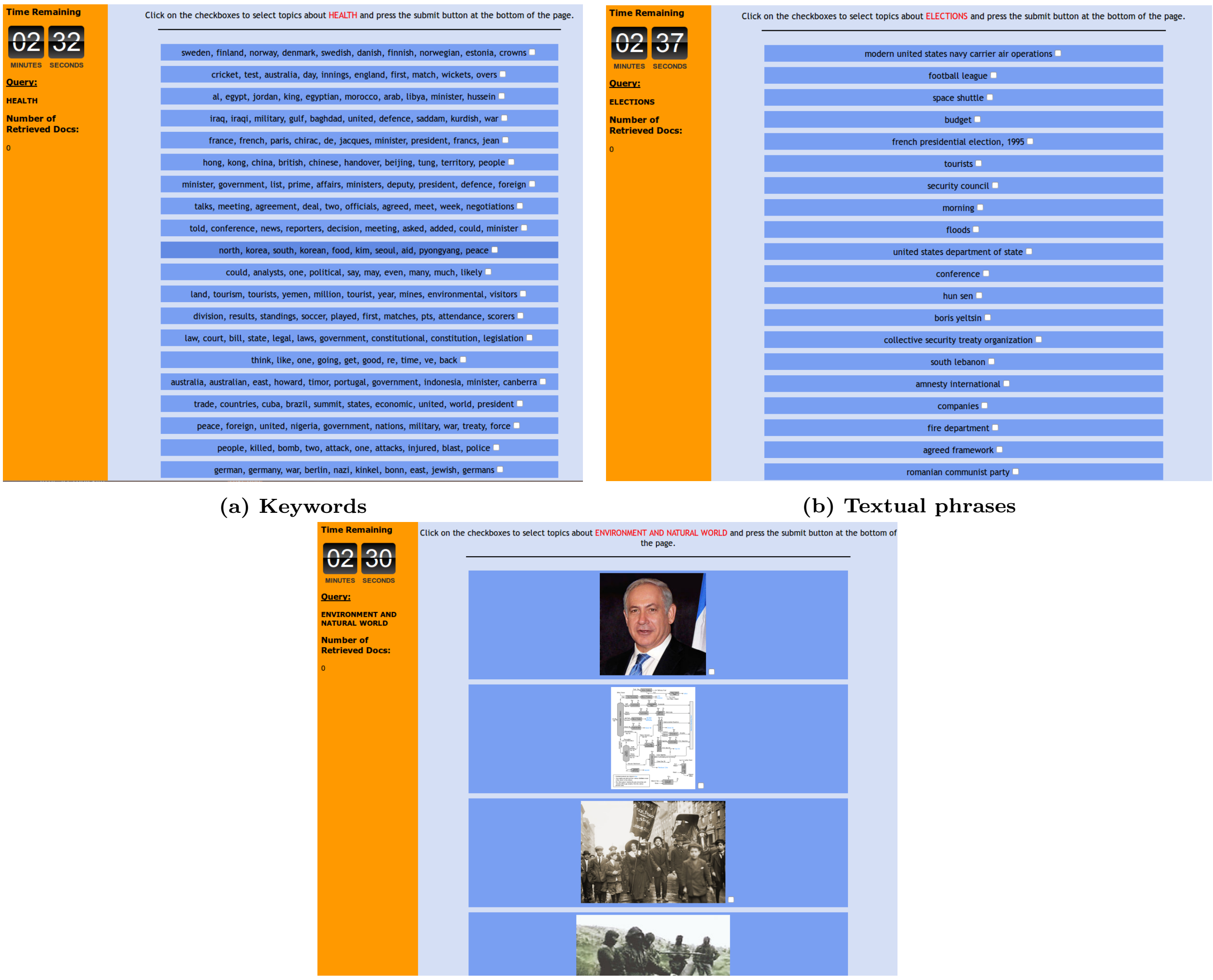
We created three browsing systems based on the TMVE. The three systems used different ways of representing topics: (1) keywords, (2) textual phrases, and (3) images. The keywords are created using a standard approach (see Section 3.3.1), the textual labels are generated from Wikipedia article titles [15] (see Section 3.3.2) while image labels are generated using publicly available images from Wikipedia [16] (see Section 3.3.3). By default, the TMVE only supports keyword representation of topics, therefore we modified it to support textual and image labels. Table 2 shows examples of the labels generated by the three approaches for a sample topic.

In addition, in the topic page, each topic is associated with its top-300 most probable documents within the topic. We restrict the number of documents shown to the user for each topic to avoid the task becoming overwhelming.

3.3.1 Keywords

Keywords are generated using the default approach of TMVE, i.e. selecting the 10 keywords with the highest marginal probabilities for the topic. This is the standard approach to representing topics used within the topic modelling research community.

[2] `https://www.cs.princeton.edu/~blei/lda-c/index.html`



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) Keywords

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 phrases

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.

3.3.2 Textual Labels

Textual labels are generated using a previously-proposed approach [15]. The labels of a topic are generated in two phases: candidate generation and candidate ranking.

In candidate generation, we use the top-7 topic terms to search Wikipedia using Wikipedia's native search API and Google's site-restricted search. We collect the top-8 article titles returned from both search engines; 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 using the RACO lexical association method [20]. The component n-grams that pass the RACO filter constitute the secondary candidates. Lastly, we also include the top-5 topic terms as part of the 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, Student's t-test, Pearson's χ^2 test, log likelihood ratio and two conditional probability variants. Term co-occurrence for computing these measures are sampled by parsing the full collection of English Wikipedia with a sliding window of length 20 words. We also include two lexical properties of the candidate as features. We combine all the features using a support vector regression model to rank the candidates.[3] The highest ranked candidate is selected as the textual label for the topic.

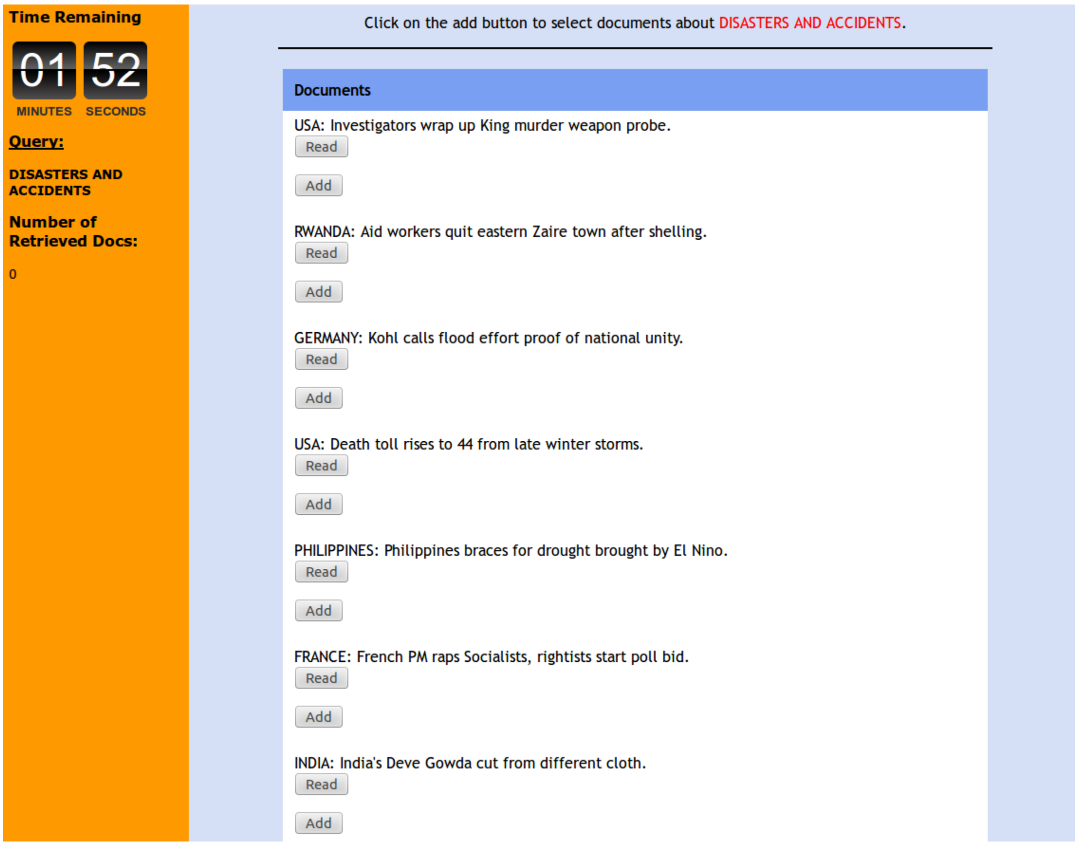
3.3.3 Image Labels

We associate topics with image labels using an existing approach [16]. 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[4]. The search is restricted to English Wikipedia[5] 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.

[3] The model is trained using the annotation collected by the authors in [15].

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

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



Time Remaining

01 52

Query:

DISASTERS AND ACCIDENTS

Number of Retrieved Docs:

0

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

Documents

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.

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 tokenisation and removal of stop words.

Visual information is extracted using low-level image keypoint descriptors, i.e. SIFT features [32, 33] sensitive to colour information. Image features are extracted using dense sampling and described using Opponent colour SIFT descriptors provided by the *colordescriptor* package.[6] The SIFT features are clustered to form a visual codebook of 1,000 visual words using k-Means 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) [34] 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 using PMI as in [16]. The image with the highest PageRank score is selected as 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 relevant 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 (keywords, textual label or image), and a query. They were then asked to identify all topics that were potentially relevant to the query. Figure 1 shows the topic browser interface for the three different modalities. In the second step, the participant 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.

We also asked users to fill 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 a score from 1 to 7 to each question. First, we asked about the usefulness of the different topic representations, i.e. keywords, 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:

[6] `http://koen.me/research/colordescriptors`

- How useful were the keywords to represent topics? (Usefulness (Keywords))
- How useful were the textual phrases to represent topics? (Usefulness (Textual label))
- How useful were the images to represent 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 Universities of Sheffield, 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. 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

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. Therefore, topic representations that are difficult to interpret will require more time for participants to understand them, 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, together with the total number of documents retrieved for each modality. Representing topics using lists of keywords results in the lowest number of documents retrieved both overall $(1,086)$ and for the majority of the queries. The number of documents retrieved when topics are represented by textual labels is highest $(1,264)$, suggesting that topics represented by textual phrases are easier to interpret than the keyword representation, making topic selection faster. The number of documents retrieved for the image representation is slightly higher than keywords but lower than textual labels.

The number of retrieved documents is high for queries that are associated with many relevant documents (*Sports* in keywords, textual 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 for 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*.

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

Query	Keywords	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

We further examined the role of the queries in the number of retrieved documents. 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 keywords and textual labels ($r = 0.76$) and keywords and image labels ($r = 0.74$), while the correlation between textual 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. However, the time required to interpret topic representations has a direct impact on the number of retrieved documents. For example, there is an overlap between the top-5 and bottom-5 queries in terms of the number of retrieved documents. In addition, we observed that the correlation between keywords and textual labels, and keywords and image labels is higher than

Table 4: Precision for each query and topic representation.

Query	Keywords	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

the correlation between textual and image labels. The main reason might be that both textual and image labels are automatically generated, which introduces noise. Comparing two noisy methods has 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 five users. Results are shown in Table 4. Keywords achieve a higher precision (0.59) than either textual (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 and image labels are assigned using automatic 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 keywords and textual labels ($r = 0.83$), and keywords and image labels ($r = 0.84$). Correlation between textual 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 mass 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 mass 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 probability mass for relevant and irrelevant documents for a given query is computed as follows:

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

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

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 mass 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 keyword representation. Textual labels perform best, while image labels obtain comparable performance. This confirms our intuition that labels can be interpreted faster than the sets of keywords. 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 mass for irrelevant topics selected using the labelling algorithms is higher than keywords. Using sets of keywords, participants select a lower number of irrelevant topics, which results to lower irrelevant probability mass. The main reason might be the false labels assigned to topics by these algorithms resulting in irrelevant topic selection by users.

4.4 Post-task Questionnaire

The main finding of the post-task questionnaire is that all of the modalities achieve similar scores in usefulness. Keywords achieve the highest score (4.33) while textual labels are close behind (4.26), and image labels slightly lower again (4.00). This demonstrates that the different topic represen-

Table 5: Document relevance based on topic selection.

	Relevant Mass			Irrelevant Mass		
Query	Keywords	Text	Image	Keywords	Text	Image
Travel & Tourism	0.00436	0.03653	0.00152	0.00034	0.02589	0.03924
Domestic Politics (USA)	0.29437	0.03453	0.09991	0.04192	0.09427	0.00013
War - Civil War	0.03034	0.00093	0.15449	0.06605	0.00026	0.02648
Biographies, Personalities, People	0.00008	0.00015	0.00014	0.04188	0.04474	0.03771
Defence	0.00032	0.00561	0.00006	0.00055	0.05458	0.00134
Crime, Law Enforcement	0.00761	0.04629	0.00019	0.00704	0.17814	0.00002
Religion	0.17583	0.02831	0.01108	0.09557	0.00062	0.05649
Disasters & Accidents	0.34822	0.09963	0.26145	0.03992	0.01217	0.03145
International Relations	0.0406	0.11082	0.01295	0.04091	0.01943	0.18198
Science & Technology	0.03895	0.21093	0.06506	0.06775	0.00027	0.01576
Employment/Labour	0.05519	0.17058	0.28647	0.00258	0.00064	0.00043
Government Finance	0.00146	0.43043	0.09921	0.0201	0.16189	0.22915
Weather	0.37622	0.88419	0.33411	0.10483	0.26126	0.00106
Elections	0.25321	0.05976	0.13636	0.03687	0.03721	0.02605
Environment & Natural World	0.07438	0.59104	0.04608	0.02711	0.1911	0.04151
Arts, Culture, Entertainment	0.0145	0.00008	0.00039	0.0252	0.32587	0.0036
Health	0.00324	0.12087	0.00126	0.00617	0.19613	0.02657
European Commission (EC) Institutions	0.00076	0.08629	0.00485	0.06255	0.00087	0.00022
Sports	0.08301	0.2506	1.38126	0.00002	0.00905	0.07218
Welfare, Social Services	0.02793	0	0.00005	0.11014	0.21971	0.36088
Average	0.09	0.16	0.15	0.04	0.09	0.06

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

The average score for Query Familiarity (4.40) denotes that the majority of the users were quite familiar with the semantic content of the queries. It is unlikely that users were unable to find relevant documents because they were unfamiliar with the queries.

Finally, we observed that participants found the retrieval task quite challenging (3.53). This may reflect the nature of the task and the limited time available for each query.

5. CONCLUSION

We compared different representations for automatically-generated topics within an exploratory browsing interface. The representations were: (1) lists of keywords, (2) textual labels, and (3) image labels. Three versions of the search interface were created, each using a different topic representation. An experiment was carried out in which users were asked to retrieve relevant documents using the interface.

Results show that participants are able to identify more documents when labels are used to represent topics, than when keywords are used. This demonstrates that the labels are a useful way of summarising the content of the topics, giving users more time to identify documents for each query and more time to explore the collection.

A greater proportion of the retrieved documents are relevant to the query for keywords than either type of label. This suggests that the keywords contain more accurate information than the labels, which is to be expected since the labels are effectively summaries of the topics and, since they are generated automatically, inevitably contain some errors [15, 16]. Despite this the number of relevant documents retrieved is very similar for all approaches.

Results indicate that automatically generated labels are a promising approach for representing topics within search interfaces. They have the advantage of being more compact than the lists of keywords that are normally used which provides more flexibility in the creation of interfaces. Retrieval performance is comparable to when keywords 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.

6. REFERENCES

[1] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.

[2] M. A. Hearst, *Search User Interfaces.* Cambridge, UK: Cambridge University Press, 2009.

[3] C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proc. of IEEE Sympos. on Visual Analytics Sci. and Technology (VAST 2009).* IEEE, 2009, pp. 91–98.

[4] M. A. Hearst, "Clustering versus faceted categories for information exploration," *Commun. of the ACM*, vol. 49, no. 4, pp. 59–61, 2006.

[5] G. Smith, M. Czerwinski, B. R. Meyers, G. Robertson, and D. Tan, "Facetmap: A scalable search and browse visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 5, pp. 797–804, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. of Mach. Learning Research*, vol. 3, pp. 993–1022, 2003.

[7] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of the 22nd Annu. Int. ACM SIGIR Conf. on Research and Develop. in Inform. Retrieval (SIGIR '99)*, Berkeley, California, United States, 1999, pp. 50–57.

[8] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models," in *Proc. of the 6th Int. AAAI Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 419–422.

[9] D. Ganguly, M. Ganguly, J. Leveling, and G. J. Jones, "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)*, Dublin, Ireland, 2013, pp. 1103–1104.

[10] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "TopicNets: Visual analysis of large text corpora with topic modeling," *ACM Trans. on Intelligent Syst. Technology*, vol. 3, no. 2, pp. 23:1–23:26, 2012.

[11] A. Hinneburg, R. Preiss, and R. Schröder, "TopicExplorer: Exploring document collections with topic models," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Comput. Sci., P. A. Flach, T. Bie, and N. Cristianini, Eds. Heidelberg, Germany: Springer, 2012, vol. 7524, pp. 838–841.

[12] J. Snyder, R. Knowles, M. Dredze, M. Gormley, and T. Wolfe, "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*, Atlanta, Georgia, 2013, pp. 5–9.

[13] Q. Mei, X. Shen, and C. X. Zhai, "Automatic Labeling of Multinomial Topic Models," in *Proc. of the 13th ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD '07)*, San Jose, California, USA, 2007, pp. 490–499.

[14] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best topic word selection for topic labelling," in *Proc. of the 23rd Int. Conf. on Computational Linguistics (COLING '10)*, Beijing, China, 2010, pp. 605–613.

[15] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proc. of the 49th Annu. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011, pp. 1536–1545.

[16] N. Aletras and M. Stevenson, "Representing topics using images," in *Proc. of the 2013 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, 2013, pp. 158–167.

[17] Q. Mei and C. Zhai, "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)*, Chicago, Illinois, USA, 2005, pp. 198–207.

[18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *J. of the American Statistical Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[19] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, "Automatic Labeling of Topics," in *Proc. of the 9th Int. Conf. on Intelligent Systems Design and Applications (ICSDA '09)*, Pisa, Italy, 2009, pp. 1227–1232.

[20] K. Grieser, T. Baldwin, F. Bohnert, and L. Sonenberg, "Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness," *J. on Computing and Cultural Heritage (JOCCH)*, vol. 3, no. 3, pp. 10:1–10:20, 2011.

[21] X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, and X. Li, "Automatic labeling hierarchical topics," in *Proc. of the 21st ACM Int. Conf. on Inform. and Knowledge Manage. (CIKM '12)*, Maui, Hawai, USA, 2012.

[22] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using DBpedia," in *Proc. of the 6th ACM Int. Conf. on Web Search and Data Mining (WSDM '13)*, Rome, Italy, 2013, pp. 465–474.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, Tech. Rep. 1999-66, 1999.

[24] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi, "The Topic Browser: An interactive tool for browsing topic models," in *NIPS Workshop on Challenges of Data Visualization*, Whistler, Canada, 2010.

[25] D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel, "Visualizing search results and document collections using topic maps," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, no. 2, pp. 169–175, 2010.

[26] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "Tiara: a visual exploratory text analytic system," in *Proc. of the 16th*

ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Washington DC, USA, 2010, pp. 153–162.

[27] J. Chuang, D. Ramage, C. Manning, and J. Heer, "Interpretation and trust: designing model-driven visualizations for text analysis," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Syst.* ACM, 2012, pp. 443–452.

[28] T. Rose, M. Stevenson, and M. Whitehead, "The Reuters Corpus volume 1 – from yesterday's news to tomorrow's language resources," in *of the 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, 2002, pp. 827–832.

[29] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. of the 26th Annu. Int. ACM SIGIR Conf. on Research and Development in Inform. Retrieval (SIGIR '03)*, Toronto, Canada, 2003, pp. 127–134.

[30] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "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)*, Jeju Island, Korea, 2012, pp. 952–961.

[31] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proc. of the 10th Int. Conf. on Computational Semantics (IWCS 2013) – Long Papers*, Potsdam, Germany, 2013, pp. 13–22.

[32] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in *Proceedings of the 7th IEEE Int. Conf. on Comput. Vision*, Kerkyra, Greece, 1999, pp. 1150–1157.

[33] ——, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. of Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[34] T. Haveliwala, S. Kamvar, and G. Jeh, "An analytical comparison of approaches to personalizing PageRank," Stanford InfoLab, Tech. Rep. 2003-35, 2003.