

aletras_2017_evaluating_topic_representations_for_exploring_document_collections

Year

2017

Author(s)

Nikolaos Aletras and Timothy Baldwin and Jey Han Lau and Mark Stevenson

Title

Evaluating Topic Representations for Exploring Document Collections

Venue

Journal of the Association for Information Science and Technology

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Established approach

Underlying technique

1. label extraction from Wikipedia titles and label ranking based on multiple lexical association feature

2. Image extraction from Wikipedia and ranking using a graph-based method

Topic labeling parameters

Phrase labels

Considered topic terms: 7

Nr of retrieved Wikipedia articles: 8

Included topic items as candidates: 5

Image labels

Top images considered: 20

Label generation

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 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; 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. 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.

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.

The top-five terms from a topic are used to query Bing using its Search API. T

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 *colordescrptor* package.

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.

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	<i>report, investigation, officials, information, intelligence, former, government, documents, alleged, fbi</i>
Textual Phrase	Label <i>Federal Bureau of Investigation</i>
Image	Label 

Motivation

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.

Topic modeling

LDA

Topic modeling parameters

Nr of topics: 100

Nr. of topics

84 (Topics that are difficult to interpret were identified using the method of Aletras and Stevenson (2013a) and removed, leaving a total of 84 topics)

Label

textual phrase labels, and image labels

Label selection

Label quality evaluation

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.

Time Remaining

0232

MINUTES SECONDS

Query:

HEALTH

Number of Retrieved Docs:

0

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

sweden, finland, norway, denmark, swedish, danish, finnish, norwegian, estonia, crowns

cricket, test, australia, day, innings, england, first, match, wickets, overs

al, egypt, jordan, king, egyptian, morocco, arab, libya, minister, hussein

iraq, iraqi, military, gulf, baghdad, united, defence, saddam, kurdish, war

france, french, paris, chirac, de, jacques, minister, president, francs, jean

hong, kong, china, british, chinese, handover, beijing, tung, territory, people

minister, government, list, prime, affairs, ministers, deputy, president, defence, foreign

talks, meeting, agreement, deal, two, officials, agreed, meet, week, negotiations

told, conference, news, reporters, decision, meeting, asked, added, could, minister

north, korea, south, korean, food, kim, seoul, aid, pyongyang, peace

could, analysts, one, political, say, may, even, many, much, likely

land, tourism, tourists, yemen, million, tourist, year, mines, environmental, visitors

division, results, standings, soccer, played, first, matches, pts, attendance, scorers

law, court, bill, state, legal, laws, government, constitutional, constitution, legislation

think, like, one, going, get, good, re, time, ve, back

australia, australian, east, howard, timor, portugal, government, indonesia, minister, canberra

trade, countries, cuba, brazil, summit, states, economic, united, world, president

peace, foreign, united, nigeria, government, nations, military, war, treaty, force

people, killed, bomb, two, attack, one, attacks, injured, blast, police

german, germany, war, berlin, nazi, kinkel, bonn, east, jewish, germans

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

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 ☐

Time Remaining

0230

MINUTESSECONDS


Query:

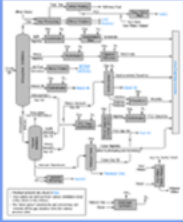
ENVIRONMENT AND
NATURAL WORLD


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.









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.

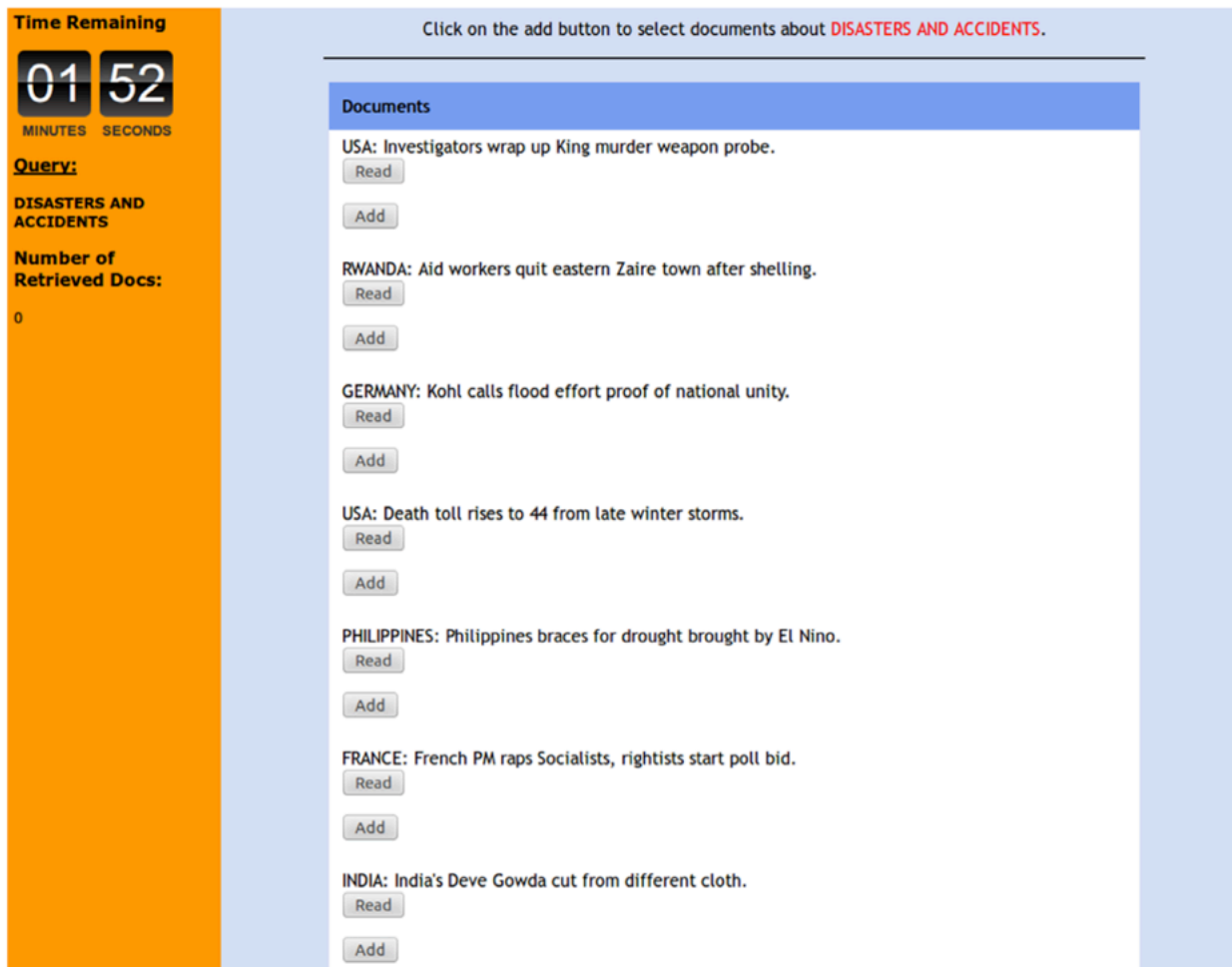


FIG. 2. Topic browsing: List of documents. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 post task questionnaire once they had completed the retrieval task. The questionnaire consisted of five questions, which were intended to provide insights into participant satisfaction with the retrieval task and the topic browsing system. Participants assigned an integer score from 1 to 7 (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”)

Results

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.

Assessors

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.

Domain

Paper: Topic labeling

Dataset: News

Problem statement

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.

Corpus

Origin: Reuters Corpus

Nr. of documents: 100,000

Details:

- Twenty topic categories were selected, and 100,000 documents were randomly extracted from the Reuters Corpus.

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

Document

Reuters news article

Pre-processing

- Tokenization
- removal of stop words
- removal of words appearing fewer than 10 times in the collection

```
@article{aletras_2017_evaluating_topic_representations_for_exploring_document_collections,
  author = {Nikolaos Aletras and Timothy Baldwin and Jey Han Lau and Mark Stevenson},
  date-added = {2023-05-03 15:06:52 +0200},
  date-modified = {2023-05-03 15:06:52 +0200},
  doi = {10.1002/asi.23574},
  journal = {Journal of the Association for Information Science and
```

```
Technology},  
  month = {jul},  
  number = {1},  
  pages = {154--167},  
  publisher = {Wiley},  
  title = {Evaluating topic representations for exploring document  
collections},  
  url = {https://doi.org/10.1002%2Fasi.23574},  
  volume = {68},  
  year = 2017}
```

#Thesis/Papers/BS