Labelling Topics using Unsupervised Graph-based Methods

Nikolaos Aletras and **Mark Stevenson**
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP
United Kingdom
{n.aletras, m.stevenson}@dcs.shef.ac.uk

Abstract

This paper introduces an unsupervised graph-based method that selects textual labels for automatically generated topics. Our approach uses the topic keywords to query a search engine and generate a graph from the words contained in the results. PageRank is then used to weigh the words in the graph and score the candidate labels. The state-of-the-art method for this task is supervised (Lau et al., 2011). Evaluation on a standard data set shows that the performance of our approach is consistently superior to previously reported methods.

1 Introduction

Topic models (Hofmann, 1999; Blei et al., 2003) have proved to be a useful way to represent the content of document collections, e.g. (Chaney and Blei, 2012; Ganguly et al., 2013; Gretarsson et al., 2012; Hinneburg et al., 2012; Snyder et al., 2013). In these interfaces, topics need to be presented to users in an easily interpretable way. A common way to represent topics is as set of keywords generated from the n terms with the highest marginal probabilities. For example, a topic about the global financial crisis could be represented by its top 10 most probable terms: FINANCIAL, BANK, MARKET, GOVERNMENT, MORTGAGE, BAILOUT, BILLION, STREET, WALL, CRISIS. But interpreting such lists is not always straightforward, particularly since background knowledge may be required (Chang et al., 2009).

Textual labels could assist with the interpretations of topics and researchers have developed methods to generate these automatically (Mei et al., 2007; Lau et al., 2010; Lau et al., 2011). For example, a topic which has keywords SCHOOL, STUDENT, UNIVERSITY, COLLEGE, TEACHER, CLASS, EDUCATION, LEARN, HIGH, PROGRAM, could be labelled as EDUCATION and a suitable label for the topic shown above would be GLOBAL FINANCIAL CRISIS. Approaches that make use of alternative modalities, such as images (Aletras and Stevenson, 2013), have also been proposed.

Mei et al. (2007) label topics using statistically significant bigrams identified in a reference collection. Magatti et al. (2009) introduced an approach for labelling topics that relied on two hierarchical knowledge resources labelled by humans, while Lau et al. (2010) proposed selecting the most representative word from a topic as its label. Hulpus et al. (2013) make use of structured data from DBpedia to label topics.

Lau et al. (2011) proposed a method for automatically labelling topics using information from Wikipedia. A set of candidate labels is generated from Wikipedia article titles by querying using topic terms. Additional labels are then generated by chunk parsing the article titles to identify n-grams that represent Wikipedia articles as well. Outlier labels (less relevant to the topic) are identified and removed. Finally, the top-5 topic terms are added to the candidate set. The labels are ranked using Support Vector Regression (SVR) (Vapnik, 1998) and features extracted using word association measures (i.e. PMI, t-test, χ^2 and Dice coefficient), lexical features and search engine ranking. Lau et al. (2011) report two versions of their approach, one unsupervised (which is used as a baseline) and another which is supervised. They reported that the supervised version achieves better performance than a previously reported approach (Mei et al., 2007).

This paper introduces an alternative graph-based approach which is unsupervised and less computationally intensive than Lau et al. (2011). Our method uses topic keywords to form a query. A graph is generated from the words contained in the search results and these are then ranked using the PageRank algorithm (Page et al., 1999; Mihal-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 631–636,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

```
{'Description': 'Microsoft will accelerate your journey to cloud computing with an
agile and responsive datacenter built from your existing technology investments.',
'DisplayUrl': 'www.microsoft.com/en-us/server-cloud/datacenter/virtualization.aspx',
'ID': 'a42b0908-174e-4f25-b59c-70bdf394a9da',
'Title': 'Microsoft | Server & Cloud | Datacenter | Virtualization ...',
'Url': 'http://www.microsoft.com/en-us/server-cloud/datacenter/virtualization.aspx',
... }
```

Figure 1: Sample of the metadata associated with a search result.

cea and Tarau, 2004). Evaluation on a standard data set shows that our method consistently outperforms the best performing previously reported method, which is supervised (Lau et al., 2011).

2 Methodology

We use the topic keywords to query a search engine. We assume that the search results returned are relevant to the topic and can be used to identify and weigh relevant keywords. The most important keywords can be used to generate keyphrases for labelling the topic or weight pre-existing candidate labels.

2.1 Retrieving and Processing Text Information

We use the approach described by Lau et al. (2011) to generate candidate labels from Wikipedia articles. The 10 terms with the highest marginal probabilities in the topic are used to query Wikipedia and the titles of the articles retrieved used as candidate labels. Further candidate labels are generated by processing the titles of these articles to identify noun chunks and n-grams within the noun chunks that are themselves the titles of Wikipedia articles. Outlier labels, identified using a similarity measure (Grieser et al., 2011), are removed. This method has been proved to produce labels which effectively summarise a topic's main subject.

However, it should be noted that our method is flexible and could be applied to any set of candidate labels. We have experimented with various approaches to candidate label generation but chose to report results using the approach described by Lau et al. (2011) to allow direct comparison of approaches.

Information obtained from web searches is used to identify the best labels from the set of candidates. The top *n* keywords, i.e. those with highest marginal probability within the topic, are used to form a query which was submitted to the Bing[1] search engine. Textual information included in the Title field[2] of the search results metadata was extracted. Each title was tokenised using openNLP[3] and stop words removed.

Figure 1 shows a sample of the metadata associated with a search result for the topic: VMWARE, SERVER, VIRTUAL, ORACLE, UPDATE, VIRTUALIZATION, APPLICATION, INFRASTRUCTURE, MANAGEMENT, MICROSOFT.

2.2 Creating a Text Graph

We consider any remaining words in the search result metadata as nodes, $v \in V$, in a graph $G = (V, E)$. Each node is connected to its neighbouring words in a context window of $\pm n$ words. In the previous example, the words added to the graph from the Title of the search result are *microsoft, server, cloud, datacenter* and *virtualization*.

We consider both unweighted and weighted graphs. When the graph is unweighted we assume that all the edges have a weight $e = 1$. In addition, we weight the edges of the graph by computing the relatedness between two nodes, v_i and v_j, as their normalised Pointwise Mutual Information (NPMI) (Bouma, 2009). Word co-occurrences are computed using Wikipedia as a a reference corpus. Pairs of words are connected with edges only if NPMI$(w_i, w_j) > 0.2$ avoiding connections between words co-occurring by chance and hence introducing noise.

2.3 Identifying Important Terms

Important terms are identified by applying the PageRank algorithm (Page et al., 1999) in a similar way to the approach used by Mihalcea and

[1] http://www.bing.com/

[2] We also experimented with using the Description field but found that this reduced performance.

[3] http://opennlp.apache.org/

Tarau (2004) for document keyphrase extraction. The PageRank score (Pr) over G for a word (v_i) can be computed by the following equation:

$$Pr(v_i) = d \cdot \sum_{v_j \in C(v_i)} \frac{sim(v_i, v_j)}{\sum\limits_{v_k \in C(v_j)} sim(v_j, v_k)} Pr(v_j) + (1 - d)\mathbf{v} \quad (1)$$

where $C(v_i)$ denotes the set of vertices which are connected to the vertex v_i. d is the damping factor which is set to the default value of $d = 0.85$ (Page et al., 1999). In standard PageRank all elements of the vector $\mathbf{v}$ are the same, $\frac{1}{N}$ where N is the number of nodes in the graph.

2.4 Ranking Labels

Given a candidate label $L = \{w_1, ..., w_m\}$ containing m keywords, we compute the score of L by simply adding the PageRank scores of its constituent keywords:

$$Score(L) = \sum_{i=1}^{m} Pr(w_i) \quad (2)$$

The label with the highest score amongst the set of candidates is selected to represent the topic. We also experimented with normalised versions of the score, e.g. mean of the PageRank scores. However, this has a negative effect on performance since it favoured short labels of one or two words which were not sufficiently descriptive of the topics. In addition, we expect that candidate labels containing words that do not appear in the graph (with the exception of stop words) are unlikely to be good labels for the topic. In these cases the score of the candidate label is set to 0. We also experimented with removing this restriction but found that it lowered performance.

3 Experimental Evaluation

3.1 Data

We evaluate our method on the publicly available data set published by Lau et al. (2011). The data set consists of 228 topics generated using text documents from four domains, i.e. blog posts (**BLOGS**), books (**BOOKS**), news articles (**NEWS**) and scientific articles from the biomedical domain (**PUBMED**). Each topic is represented by its ten most probable keywords. It is also associated with candidate labels and human ratings denoting the appropriateness of a label given the topic. The full data set consists of approximately 6,000 candidate labels (27 labels per topic).

3.2 Evaluation Metrics

Our evaluation follows the framework proposed by Lau et al. (2011) using two metrics, i.e. **Top-1 average rating** and **nDCG**, to compare various labelling methods.

Top-1 average rating is the average human rating (between 0 and 3) assigned to the top-ranked label proposed by the system. This provides an indication of the overall quality of the label the system judges as the best one.

Normalised discounted cumulative gain (**nDCG**) (Järvelin and Kekäläinen, 2002; Croft et al., 2009) compares the label ranking proposed by the system to the ranking provided by human annotators. The discounted cumulative gain at position p, DCG_p, is computed using the following equation:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)} \quad (3)$$

where rel_i is the relevance of the label to the topic in position i. Then nDCG is computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (4)$$

where $IDCG_p$ is the superviseed ranking of the image labels, in our experiments this is the ranking provided by the scores in the human annotated data set.

3.3 Model Parameters

Our proposed model requires two parameters to be set: the context window size when connecting neighbouring words in the graph and the number of the search results considered when constructing the graph.

We experimented with different sizes of context window, n, between ±1 words to the left and right and all words in the title. The best results were obtained when $n = 2$ for all of the domains. In addition, we experimented with varying the number of search results between 10 and 300. We observed no noticeable difference in the performance when the number of search results is equal or greater than 30 (see below). We choose to report results obtained using 30 search results for each topic. Including more results did not improve performance but required additional processing.

Domain	Model	Top-1 Av. Rating	nDCG-1	nDCG-3	nDCG-5
BLOGS	Lau et al. (2011)-U	1.84	0.75	0.77	0.79
	Lau et al. (2011)-S	1.98	0.81	0.82	0.83
	PR	2.05†	0.83	0.84	0.83
	PR-NPMI	2.08†	0.84	0.84	0.83
	Upper bound	2.45	1.00	1.00	1.00
BOOKS	Lau et al. (2011)-U	1.75	0.77	0.77	0.79
	Lau et al. (2011)-S	1.91	0.84	0.81	0.83
	PR	1.98†	0.86	0.88	0.87
	PR-NPMI	2.01†	0.87	0.88	0.87
	Upper bound	2.29	1.00	1.00	1.00
NEWS	Lau et al. (2011)-U	1.96	0.80	0.79	0.78
	Lau et al. (2011)-S	2.02	0.82	0.82	0.84
	PR	2.04†	0.83	0.81	0.81
	PR-NPMI	2.05†	0.83	0.81	0.81
	Upper bound	2.45	1.00	1.00	1.00
PUBMED	Lau et al. (2011)-U	1.73	0.75	0.77	0.79
	Lau et al. (2011)-S	1.79	0.77	0.82	0.84
	PR	1.88†‡	0.80	0.80	0.80
	PR-NPMI	1.90†‡	0.81	0.80	0.80
	Upper bound	2.31	1.00	1.00	1.00

Table 1: Results for Various Approaches to Topic Labelling (†: significant difference (t-test, $p < 0.05$) to Lau et al. (2011)-U; ‡: significant difference ($p < 0.05$) to Lau et al. (2011)-S).

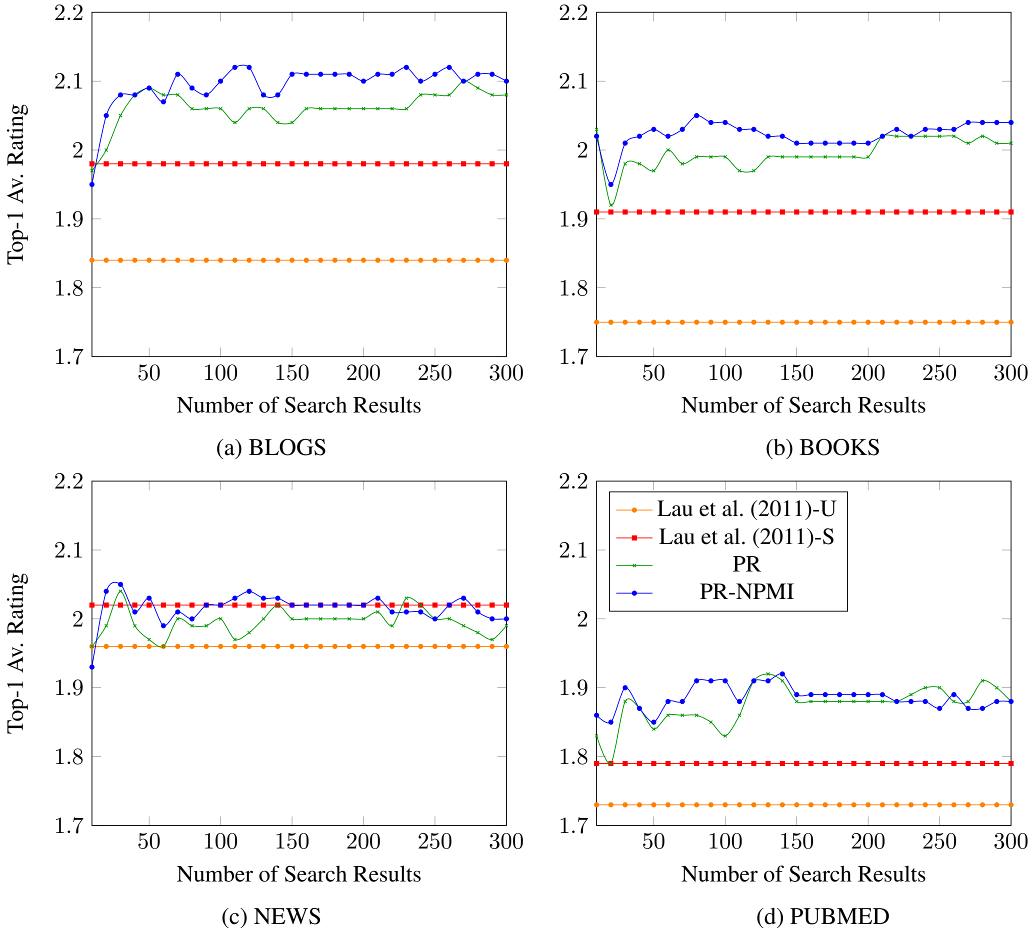
4 Results and Discussion

Results are shown in Table 1. Performance when PageRank is applied to the unweighted (**PR**) and NPMI-weighted graphs (**PR-NPMI**) (see Section 2.2) is shown. Performance of the best unsupervised (**Lau et al. (2011)-U**) and supervised (**Lau et al. (2011)-S**) methods reported by Lau et al. (2011) are shown. Lau et al. (2011)-U uses the average χ^2 scores between the topic keywords and the label keywords while Lau et al. (2011)-S uses SVR to combine evidence from all features. In addition, upper bound figures, the maximum possible value given the scores assigned by the annotators, are also shown.

The results obtained by applying PageRank over the unweighted graph (2.05, 1.98, 2.04 and 1.88) are consistently better than the supervised and unsupervised methods reported by Lau et al. (2011) for the Top-1 Average scores and this improvement is observed in all domains. The difference is significant (t-test, $p < 0.05$) for the unsupervised method. A slight improvement in performance is observed when the weighted graph is used (2.08, 2.01, 2.05 and 1.90). This is expected since the weighted graph contains additional information about word relatedness. For example, the word *hardware* is more related and, therefore, closer in the graph to the word *virtualization* than to the word *investments*.

Results from the nDCG metric imply that our methods provide better rankings of the candidate labels in the majority of the cases. It is outperformed by the best supervised approach in two domains, NEWS and PUBMED, using the nDCG-3 and nDCG-5 metrics. However, the best label proposed by our methods is judged to be better (as shown by the nDCG-1 and Top-1 Av. Rating scores), demonstrating that it is only the lower ranked labels in our approach that are not as good as the supervised approach.

An interesting finding is that, although limited in length, the textual information in the search result's metadata contain enough salient terms relevant to the topic to provide reliable estimates of



(a) BLOGS

(b) BOOKS

(c) NEWS

(d) PUBMED

Figure 2: Top-1 Average Rating obtained for different number of search results.

term importance. Consequently, it is not necessary to measure semantic similarity between topic keywords and candidate labels as previous approaches have done. In addition, performance improvement gained from using the weighted graph is modest, suggesting that the computation of association scores over a large reference corpus could be omitted if resources are limited.

In Figure 2, we show the scores of Top-1 average rating obtained in the different domains by experimenting with the number of search results used to generate the text graph. The most interesting finding is that performance is stable when 30 or more search results are considered. In addition, we observe that quality of the topic labels in the four domains remains stable, and higher than the supervised method, when the number of search results used is between 150 and 200. The only domain in which performance of the supervised method is sometimes better than the approach proposed here is NEWS. The main reason is that news topics are more fine grained and the candidate labels of better quality (Lau et al., 2011) which has direct impact in good performance of ranking methods.

5 Conclusion

We described an unsupervised graph-based method to associate textual labels with automatically generated topics. Our approach uses results retrieved from a search engine using the topic keywords as a query. A graph is generated from the words contained in the search results metadata and candidate labels ranked using the PageRank algorithm. Evaluation on a standard data set shows that our method consistently outperforms the supervised state-of-the-art method for the task.

Acknowledgments

We would like to thank Jey Han Lau for providing us with the labels selected by Lau et al. (2011)-U and Lau et al. (2011)-S. We also thank Daniel Preoţiuc-Pietro for his useful comments on early drafts of this paper.

References

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

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of GSCL*.

Allison June-Barlow Chaney and David M. Blei. 2012. Visualizing topic models. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland.

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

Bruce W. Croft, Donald Metzler, and Trevor Strohman. 2009. *Search engines: Information retrieval in practice*. Addison-Wesley.

Debasis Ganguly, Manisha Ganguly, Johannes Leveling, and Gareth J.F. Jones. 2013. TopicVis: A GUI for Topic-based feedback and navigation. In *Proceedings of the Thirty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 13)*, Dublin, Ireland.

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

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

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

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 50–57, Berkeley, California, United States.

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

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

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

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

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

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

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pages 404–411, Barcelona, Spain.

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

Justin Snyder, Rebecca Knowles, Mark Dredze, Matthew Gormley, and Travis Wolfe. 2013. Topic models and metadata for visualizing text corpora. In *Proceedings of the 2013 NAACL-HLT Demonstration Session*, pages 5–9, Atlanta, Georgia. Association for Computational Linguistics.

Vladimir N Vapnik. 1998. *Statistical learning theory*. Wiley, New York.