Representing Topics Using Images

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

Abstract

Topics generated automatically, e.g. using LDA, are now widely used in Computational Linguistics. Topics are normally represented as a set of keywords, often the n terms in a topic with the highest marginal probabilities. We introduce an alternative approach in which topics are represented using images. Candidate images for each topic are retrieved from the web by querying a search engine using the top n terms. The most suitable image is selected from this set using a graph-based algorithm which makes use of textual information from the metadata associated with each image and features extracted from the images themselves. We show that the proposed approach significantly outperforms several baselines and can provide images that are useful to represent a topic.

1 Introduction

Topic models are statistical methods for summarising the content of a document collection using latent variables known as topics (Hofmann, 1999; Blei et al., 2003). Within a model, each topic is a multinomial distribution over words in the collection while documents are represented as distributions over topics. Topic modelling is now widely used in Natural Language Processing (NLP) and has been applied to a range of tasks including word sense disambiguation (Boyd-Graber et al., 2007), multi-document summarisation (Haghighi and Vanderwende, 2009), information retrieval (Wei and Croft, 2006), image labelling (Feng and Lapata, 2010a) and visualisation of document collections (Chaney and Blei, 2012).

Topics are often represented by using the n terms with the highest marginal probabilities in the topic to generate a set of keywords. For example, *wine*, *bottle*, *grape*, *flavour*, *dry*. Interpreting such lists may not be straightforward, particularly since there may be no access to the source collection used to train the model. Therefore, researchers have recently begun developing automatic methods to generate meaningful and representative labels for topics. These techniques have focussed on the creation of textual labels (Mei et al., 2007; Lau et al., 2010; Lau et al., 2011).

An alternative approach is to represent a topic using an illustrative image (or set of images). Images have the advantage that they can be understood quickly and are language independent. This is particularly important for applications in which the topics are used to provide an overview of a collection with many topics being shown simultaneously (Chaney and Blei, 2012; Gretarsson et al., 2012; Hinneburg et al., 2012).

This paper explores the problem of selecting images to illustrate automatically generated topics. Our approach generates a set of candidate images for each topic by querying an image search engine with the top n topic terms. The most suitable image is selected using a graph-based method that makes use of both textual and visual information. Textual information is obtained from the metadata associated with each image while visual features are extracted from the images themselves. Our approach is evaluated using a data set created for this study that was annotated by crowdsourcing. Results of the evaluation show that the proposed method significantly

Proceedings of NAACL-HLT 2013, pages 158–167,
Atlanta, Georgia, 9–14 June 2013.

outperforms three baselines.

The contributions of this paper are as follows: (1) introduces the problem of labelling topics using images; (2) describes an approach to this problem that makes use of multimodal information to select images from a set of candidates; (3) introduces a data set to evaluate image labelling; and (4) evaluates the proposed approach using this data set.

2 Related work

In early research on topic modelling, labels were manually assigned to topics for convenient presentation of research results (Mei and Zhai, 2005; Teh et al., 2006).

The first attempt at automatically assigning labels to topics is described by Mei et al. (2007). In their approach, a set of candidate labels are extracted from a reference collection using chunking and statistically important bigrams. Then, a relevance scoring function is defined which minimises the Kullback-Leibler divergence between word distribution in a topic and word distribution in candidate labels. Candidate labels are ranked according to their relevance and the top ranked label chosen to represent the topic.

Magatti et al. (2009) introduced an approach for labelling topics that relied on two hierarchical knowledge resources labelled by humans, the Google Directory and the OpenOffice English Thesaurus. A *topics tree* is a pre-existing hierarchical structure of labelled topics. The Automatic Labelling Of Topics algorithm computes the similarity between LDA inferred topics and topics in a *topics 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. (2010) proposed selecting the most representative word from a topic as its label. A label is selected by computing the similarity between each word and all the 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 to achieve results above a baseline (the most probable word in the topic).

In more recent work, Lau et al. (2011) proposed a method for automatically labelling topics by making use of Wikipedia article titles as candidate labels. The candidate labels are ranked using information from word association measures, lexical features and an Information Retrieval technique. Results showed that this ranking method achieves better performance than a previous approach (Mei et al., 2007).

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

Hulpus et al. (2013) make use of the structured data in DBpedia[1] to label topics. Their approach maps topic words to DBpedia concepts. The best concepts are identified by applying graph centrality measures which assume that words that co-occurring in text are likely to refer to concepts that are close in the DBpedia graph.

Our own work differs from the approaches described above since, to our knowledge, it is the first to propose labelling topics with images rather than text.

Recent advances in computer vision has lead to the development of reliable techniques for exploiting information available in images (Datta et al., 2008; Szeliski, 2010) and these have been combined with NLP (Feng and Lapata, 2010a; Feng and Lapata, 2010b; Agrawal et al., 2011; Bruni et al., 2011). The closest work to our own is the text illustration techniques which have been proposed for story picturing (Joshi et al., 2006) and news articles illustration (Feng and Lapata, 2010b). The input to text illustration models is a textual document and a set of image candidates. The goal of the models is to associate the document with the correct image. Moreover, the problem of ranking images returned from a text query is related to, but different from, the one explored in our paper. Those approaches used queries that were much smaller (e.g. between one and three words) and more focussed than the ones

[1] http://dbpedia.org

we use (Jing and Baluja, 2008). In our work, the input is a topic and the aim is to associate it with an image, or images, denoting the main thematic subject.

3 Labelling Topics

In this section we propose an approach to identifying images to illustrate automatically generated topics. It is assumed that there are no candidate images available so the first step (Section 3.1) is to generate a set of candidate images. However, when a candidate set is available the first step can be skipped.

3.1 Selecting Candidate Images

For the experiments presented here we restrict ourselves to using images from Wikipedia available under the Creative Commons licence, since this allows us to make the data available. The top-5 terms from a topic are used to query Google using its Custom Search API[2]. The search is restricted to the English Wikipedia[3] with image search enabled. The top-20 images retrieved for each search are used as candidates for the topic.

3.2 Feature Extraction

Candidate images are represented by two modalities (textual and visual) and features extracted for each.

3.2.1 Textual Information

Each image's textual information consists of the metadata retrieved by the search. The assumption here is that image's metadata is indicative of the image's content and (at least to some extent) related to the topic. The textual information is formed by concatenating the *title* and the *link* 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 tokenizing and removing stop words.

3.2.2 Visual Information

Visual information is extracted using low-level image keypoint descriptors, i.e. SIFT features (Lowe, 1999; Lowe, 2004) sensitive to colour information. SIFT features denote "interesting" areas in an image. Image features are extracted using dense sampling and described using Opponent colour SIFT descriptors provided by the *colordescriptor*[4] software. Opponent colour SIFT descriptors have been found to give the best performance in object scene and face recognition (Sande et al., 2008). 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).

3.3 Ranking Candidate Images

We rank images in the candidates set using graph-based algorithms. The graph is created by treating images as nodes and using similarity scores (textual or visual) between images to weight the edges.

3.3.1 PageRank

PageRank (Page et al., 1999) is a graph-based algorithm for identifying important nodes in a graph that was originally developed for assigning importance to web pages. It has been used for a range of NLP tasks including word sense disambiguation (Agirre and Soroa, 2009) and keyword extraction (Mihalcea and Tarau, 2004).

Let $G = (V, E)$ be a graph with a set of vertices, V, denoting image candidates and a set of edges, E, denoting similarity scores between two images. For example, $sim(V_i, V_j)$ indicates the similarity between images V_i and V_j. The PageRank score (Pr) over G for an image (V_i) can be computed by the following equation:

$$Pr(V_i) = d \cdot \sum_{V_j \in C(V_i)} \frac{sim(V_i, V_j)}{\sum_{V_k \in C(V_j)} sim(V_j, V_k)} Pr(V_j) + (1-d)\mathbf{v} \quad (1)$$

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] https://developers.google.com/apis-explorer/#s/customsearch/v1

[3] http://en.wikipedia.org

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

3.3.2 Personalised PageRank

Personalised PageRank (PPR) (Haveliwala et al., 2003) is a variant of the PageRank algorithm in which extra importance is assigned to certain vertices in the graph. This is achieved by adjusting the values of the vector $\mathbf{v}$ in equation 1 to prefer certain nodes. Nodes that are assigned high values in $\mathbf{v}$ are more likely to also be assigned a high PPR score. We make use of PPR to prefer images with textual information that is similar to the terms in the topic.

3.3.3 Weighting Graph Edges

Three approaches were compared for computing the values of $sim(V_i, V_j)$ in equation 1 used to weight the edges of the graph. Two of these make use of the textual information associated with each image while the final one relies on visual features.

The first approach is **Pointwise Mutual Information** (PMI). The similarity between a pair of images (vertices in the graph) is computed as the average PMI between the terms in their metadata. PMI is computed using word co-occurrence counts over Wikipedia identified using a sliding window of length 20. We also experimented with other word association measures but these did not perform as well. The PageRank over the graph weighted using PMI is denoted as $\mathbf{PR_{PMI}}$.

The second approach, **Explicit Semantic Analysis** (ESA) (Gabrilovich and Markovitch, 2007), is a knowledge-based similarity measure. ESA transforms the text from the image metadata into vectors that consist of Wikipedia article titles weighted by their relevance. The similarity score between these vectors is computed as the cosine of the angle between them. This similarity measure is used to create the graph and its PageRank is denoted as $\mathbf{PR_{ESA}}$.

The final approach uses the **visual features** extracted from the images themselves. The visual words extracted from the images are used to form feature vectors and the similarity between a pair of images computed as the cosine of the angle between them. The PageRank of the graph created using this approach is $\mathbf{PR_{vis}}$ and it is similar to the approach proposed by Jing and Baluja (2008) for associating images to text queries.

3.3.4 Initialising the Personalisation Vector

The personalisation vector (see above) is weighted using the similarity scores computed between the topic and its image candidates. Similarity is computed using PMI and ESA (see above). When PMI and ESA are used to weight the personalisation vector they compute the similarity between the top 10 terms for a topic and the textual metadata associated with each image in the set of candidates. We refer to the personalisation vectors created using PMI and ESA as **Per(PMI)** and **Per(ESA)** respectively.

Using PPR allows information about the similarity between the images' metadata and the topics themselves to be considered when identifying a suitable image label. The situation is different when PageRank is used since this only considers the similarity between the images in the candidate set.

The personalisation vector used by PPR is employed in combination with a graph created using one of the approaches described above. For example, the graph may be weighted using visual features and the personalisation vector created using PMI scores. This approach is denoted as PR_{vis}+Per(PMI).

4 Evaluation

This section discusses the experimental design for evaluating the proposed approaches to labelling topics with images. To our knowledge no data set for evaluating these approaches is currently available and consequently we developed one for this study[5]. Human judgements about the suitability of images are obtained through crowdsourcing.

4.1 Data

We created a data set of topics from two collections which cover a broad thematic range:

- **NYT** 47,229 New York Times news articles (included in the GigaWord corpus) that were published between May and December 2010.
- **WIKI** A set of Wikipedia categories randomly selected by browsing its hierarchy in a breadth-first-search manner starting from a few seed

[5]Data set can be downloaded from `http://staffwww.dcs.shef.ac.uk/people/N.Aletras/resources.html`.

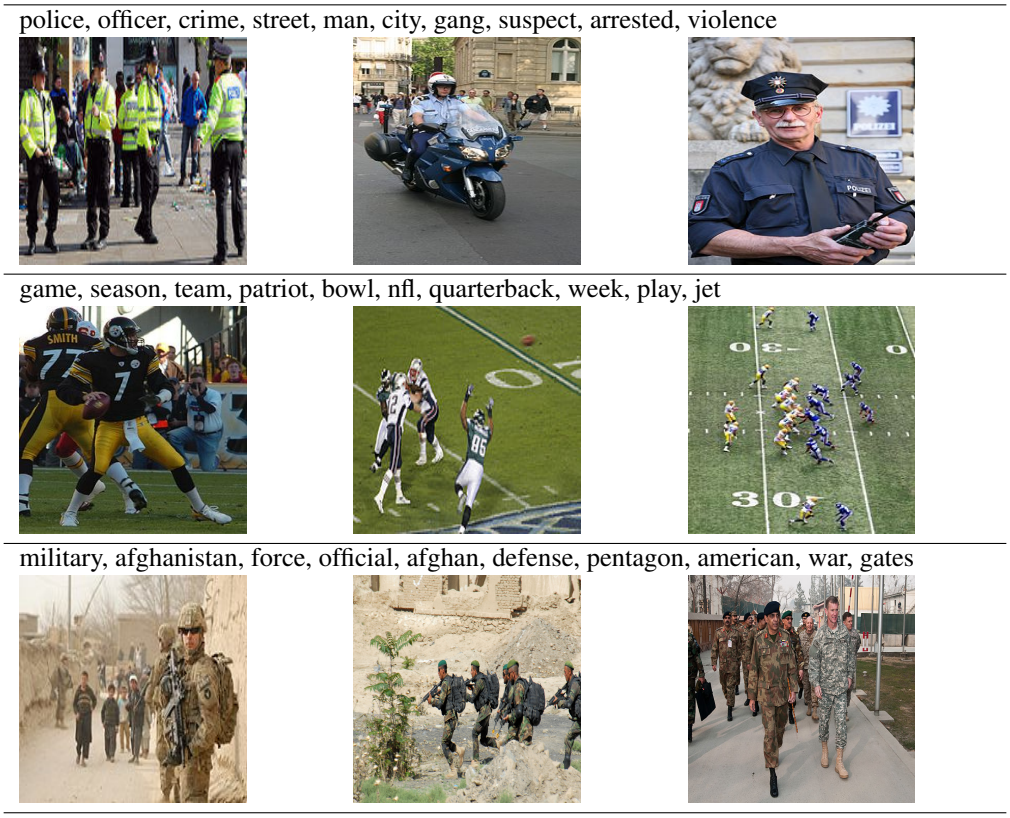

Figure 1: A sample of topics and their top-3 image candidates (i.e. with the highest average human annotations).

categories (e.g. SPORTS, POLITICS, COMPUTING). Categories that have more that 80 articles associated with them are considered. These articles are collected to produce a corpus of approximately 60,000 articles generated from 1,461 categories.

Documents in the two collections are tokenised and stop words removed. LDA was applied to learn 200 topics from NYT and 400 topics from WIKI. The *gensim* package[6] was used to implement and compute LDA. The hyperparameters (α, β) were set to $\frac{1}{num\_of\_topics}$. Incoherent topics are filtered out by applying the method proposed by Aletras and Stevenson (2013).

We randomly selected 100 topics from NYT and 200 topics from WIKI resulting in a data set of 300 topics. Candidate images for these topics were generated using the approach described in Section 3.1, producing a total of 6,000 candidate images (20 for each topic).

4.2 Human Judgements of Image Relevance

Human judgements of the suitability of each image were obtained using an online crowdsourcing platform, Crowdflower[7]. Annotators were provided with a topic (represented as a set of 10 keywords) and a candidate image. They were asked to judge how appropriate the image was as a representation of the main subject of the topic and provide a rating on a scale of 0 (completely unsuitable) to 3 (very suitable).

Quality control is important in crowdscourcing experiments to ensure reliability (Kazai, 2011). To avoid random answers, control questions with obvious answer were included in the survey. Annotations by participants that failed to answer these questions correctly or participants that gave the same rating for all pairs were removed.

[6] http://pypi.python.org/pypi/gensim

[7] http://crowdflower.com

The total number of filtered responses obtained was $62,221$ from 273 participants. Each topic-image pair was rated at least by 10 subjects. The average response for each pair was calculated in order to create the final similarity judgement for use as a gold-standard. The average variance across judges (excluding control questions) is 0.88.

Inter-Annotator agreement (IAA) is computed as the average Spearman's ρ between the ratings given by an annotator and the average ratings given by all other annotators. The average IAA across all topics was 0.50 which indicates the difficulty of the task, even for humans.

Figure 1 shows three example topics from the data set together with the images that received the highest average score from the annotators.

4.3 Evaluation Metrics

Evaluation of the topic labelling methods is carried out using a similar approach to the framework proposed by Lau et al. (2011) for labelling topics using textual labels.

Top-1 average rating is the average human rating assigned to the top-ranked label proposed by the system. This provides an indication of the overall quality of the image the system judges as the best one. The highest possible score averaged across all topics is 2.68, since for many topics the average score obtained from the human judgements is lower than 3.

The second evaluation measure is the normalized discounted cumulative gain (**nDCG**) (Järvelin and Kekäläinen, 2002; Croft et al., 2009) which compares the label ranking proposed by the system to the optimal ranking provided by humans. The discounted cumulative gain at position p (DCG_p) is computed using the following equation:

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

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

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (3)$$

where $IDCG_p$ is the optimal ranking of the image labels, in our experiments this is the ranking provided by the scores in the human annotated data set. We follow Lau et al. (2011) in computing **nDCG-1**, **nDCG-3** and **nDCG-5** for the top 1, 3 and 5 ranked system image labels respectively.

4.4 Baselines

Since there are no previous methods for labelling topics using images, we compare our proposed models against three baselines.

The **Random** baseline randomly selects a label for the topic from the 20 image candidates. The process is repeated 10,000 times and the average score of the selected labels is computed for each topic.

The more informed **Word Overlap** baseline selects the image that is most similar to the topic terms by applying a Lesk-style algorithm (Lesk, 1986) to compare metadata for each image against the topic terms. It is defined as the number of common terms between a topic and image candidate normalised by the total number of terms in the topic and image's metadata.

We also compared our approach with the ranking returned by the **Google Image Search** for the top-20 images for a specific topic.

4.5 User Study

A user study was conducted to estimate human performance on the image selection task. Three annotators were recruited and asked to select the best image for each of the 300 topics in the data set. The annotators were provided with the topic (in the form of a set of keywords) and shown all candidate images for that topic before being asked to select exactly one. The Average Top-1 Rating was computed for each annotator and the mean of these values was 2.24.

5 Results

Table 1 presents the results obtained for each of the methods on the collection of 300 topics. Results are shown for both Top-1 Average rating and nDCG.

We begin by discussing the results obtained using the standard PageRank algorithm applied to graphs weighted using PMI, ESA and visual features (PR_{PMI}, PR_{ESA} and PR_{vis} respectively). Results using PMI consistently outperform all baselines and those obtained using ESA. This suggests that distributional word association measures are more suitable for identifying useful images than knowledge-based similarity measures. The best results using

Model	Top-1 Av. Rating	nDCG-1	nDCG-3	nDCG-5
Baselines				
Random	1.79	-	-	-
Word Overlap	1.85	0.69	0.72	0.74
Google Image Search	1.89	0.73	0.75	0.77
PageRank				
PR_{PMI}	1.87	0.70	0.73	0.75
PR_{ESA}	1.81	0.67	0.68	0.70
PR_{vis}	1.96	0.73	0.75	0.76
Personalised PageRank				
PR_{PMI}+Per(PMI)	1.98	0.74	0.76	0.77
PR_{PMI}+Per(ESA)	1.92	0.70	0.72	0.74
PR_{ESA}+Per(PMI)	1.91	0.70	0.72	0.73
PR_{ESA}+Per(ESA)	1.88	0.69	0.72	0.74
PR_{vis}+Per(PMI)	2.00	0.74	0.75	0.76
PR_{vis}+Per(ESA)	1.94	0.72	0.75	0.76
User Study	2.24	–	–	–

Table 1: Results for various approaches to topic labelling.

standard PageRank are obtained when the visual similarity measures are used to weight the graph, with performance that significantly outperforms the word overlap baseline (paired t-test, $p < 0.05$). This demonstrates that visual features are a useful source of information for deciding which images are suitable topic labels.

The Personalised version of PageRank produces consistently higher results compared to standard PageRank, demonstrating that the additional information provided by comparing the image metadata with the topics is useful for this task. The best results are obtained when the personalisation vector is weighted using PMI (i.e. Per(PMI)). The best overall result for the top-1 average rating (2.00) is obtained when the graph is weighted using visual features and the personalisation vector using the PMI scores (PR_{vis}+Per(PMI)) while the best results for the various DCG metrics are produced when both the graph and the personalisation vector are weighted using PMI scores (PR_{PMI}+Per(PMI)). In addition, these two methods, PR_{vis}+Per(PMI) and PR_{PMI}+Per(PMI), perform significantly better than the word overlap and the Google Image Search baselines ($p < 0.01$ and $p < 0.05$ respectively). Weighting the personalisation vector using ESA consistently produces lower performance compared to PMI. These results indicate that graph-based methods for ranking images are useful for illustrating topics.

6 Discussion

Figure 2 shows a sample of three topics together with the top-3 candidates (left-to-right) selected by applying the PR_{vis}+Per(PMI) approach. Reasonable labels have been selected for the first two topics. On the other hand, the images selected for the third topic do not seem to be as appropriate.

We observed that inappropriate labels can be generated for two reasons. Firstly, the topic may be abstract and difficult to illustrate. For example, one of the topics in our data set refers to the subject ALGEBRAIC NUMBER THEORY and contains the terms *number, ideal, group, field, theory, algebraic, class, ring, prime, theorem*. It is difficult to find a representative image for topics such as this one. Secondly, there are topics for which none of the candidate images returned by the search engine is relevant. An example of a topic like this in our data set is one that refers to PLANTS and contains the terms *family, sources, plants, familia, order, plant, species, taxonomy, classification, genera*. The images returned by the search engine include pictures of the Sagrada Familia cathedral in Barcelona, a car called "Familia"

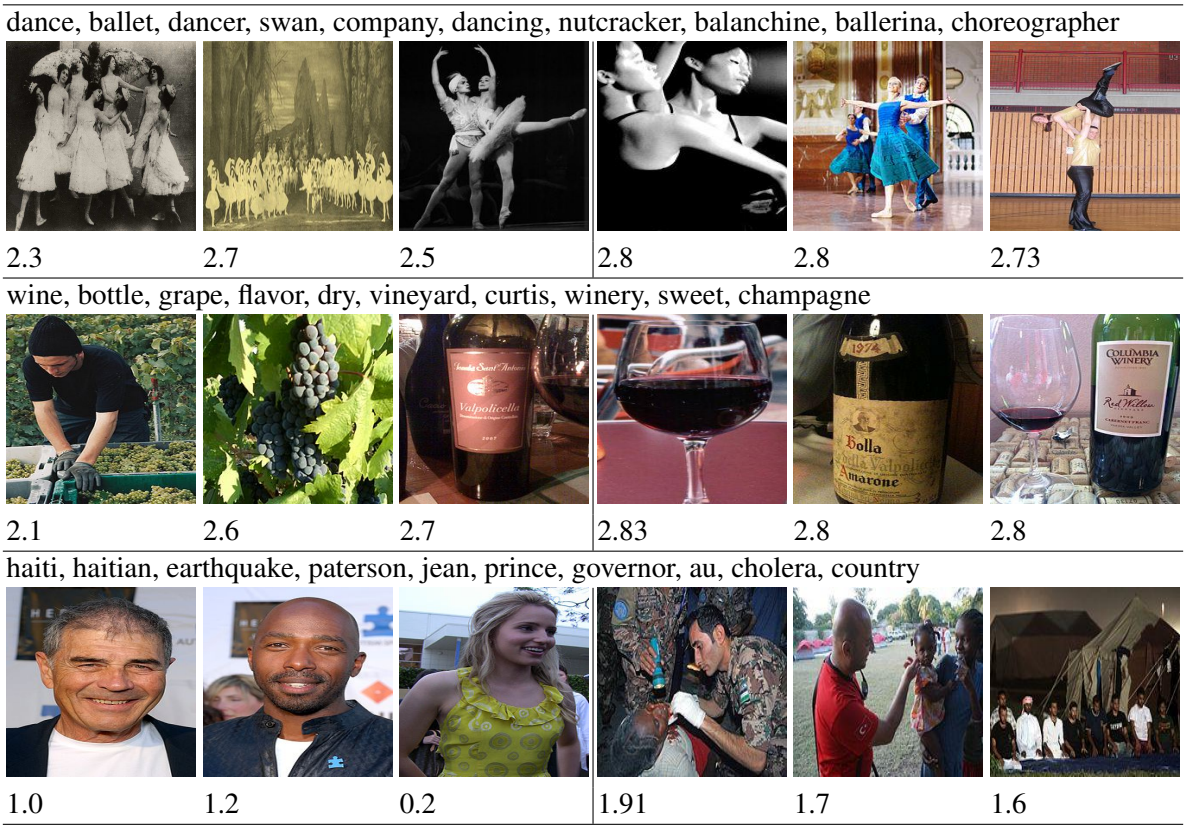

Figure 2: A sample of topics and their top-3 images selected by applying the the PR_{vis}+Per(PMI) approach (left side) and the ones with the highest average human annotations (right side). The number under each image represents its average human annotations score.

and pictures of families but no pictures of plants.

7 Conclusions

This paper explores the use of images to represent automatically generated topics. An approach to selecting appropriate images was described. This begins by identifying a set of candidate images using a search engine and then attempts to select the most suitable. Images are ranked using a graph-based method that makes use of both textual and visual information. Evaluation is carried out on a data set created for this study. The results show that the visual features are a useful source of information for this task while the proposed graph-based method significantly outperforms several baselines.

This paper demonstrates that it is possible to identify images to illustrate topics. A possible application for this technique is to represent the contents of large document collections in a way that supports rapid interpretation and can be used to enable navigation (Chaney and Blei, 2012; Gretarsson et al., 2012; Hinneburg et al., 2012). We plan to explore this possibility in future work. Other possible extensions to this work include exploring alternative approaches to generating candidate images and developing techniques to automatically identify abstract topics for which suitable images are unlikely to be found, thereby avoiding the problem cases described in Section 6.

Acknowledgments

The research leading to these results was carried out as part of the PATHS project (`http://paths-project.eu`) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pages 33–41, Athens, Greece.

Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. 2011. Enriching textbooks with images. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*, pages 1847–1856, Glasgow, Scotland, UK.

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

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

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pages 1024–1033, Prague, Czech Republic.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics (GEMS '11)*, pages 22–32, Edinburgh, UK.

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.

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

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2):1–60.

Yansong Feng and Mirella Lapata. 2010a. How many words is a picture worth? Automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249, Uppsala, Sweden.

Yansong Feng and Mirella Lapata. 2010b. Topic Models for Image Annotation and Text Illustration. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '07)*, pages 1606–1611, Hyberabad, India.

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.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado.

Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. 2003. An analytical comparison of approaches to personalizing PageRank. Technical Report 2003-35, Stanford InfoLab.

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.

Yushi Jing and Shumeet Baluja. 2008. PageRank for product image search. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, pages 307–316, Beijing, China.

Dhiraj Joshi, James Z. Wang, and Jia Li. 2006. The Story Picturing Engine—A system for automatic text illustration. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):68–89.

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

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for

topic labelling. In *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.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86)*, pages 24–26, Toronto, Ontario, Canada.

David G. Lowe. 1999. Object Recognition from Local Scale-invariant Features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157, Kerkyra, Greece.

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

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.

Xian-Li Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. 2012. Automatic labeling hierarchical topics. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, Sheraton, Maui Hawai.

Qiaozhu Mei and ChengXiang Zhai. 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)*, pages 198–207, Chicago, Illinois, USA.

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.

Koen E.A. Sande, Theo Gevers, and Cees G. M. Snoek. 2008. Evaluation of Color Descriptors for Object and Scene Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pages 1–8, Anchorage, Alaska, USA.

Richard Szeliski. 2010. *Computer Vision: Algorithms and Applications*. Springer-Verlag Inc.

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

Xing Wei and W. Bruce Croft. 2006. LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '06)*, pages 178–185, Seattle, Washington, USA.