

Interpreting Document Collections with Topic Models



A Thesis submitted to the University of Sheffield
for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Nikolaos Aletras

Department of Computer Science

The University of Sheffield

September 2014

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my supervisor, Mark Stevenson. It was a good fun working together. Mark provided me with invaluable advice and help throughout my PhD from high-level ideas to small text polishing issues. Mark also helped me comprehend the right research practice making these three years of study a real pleasure.

I am very grateful to my PhD panel committee members, Kalina Bontcheva and Steve Maddock, for their invaluable advice at the early stages of my studies.

Many thanks to my collaborators in various research projects: Trevor Cohn, Bill Lampos and Daniel Preoțiuc-Pietro in social media analysis; Eneko Agirre, Paul Clough, Samuel Fernando, Aitor Gonzalez-Agirre, Mark Hall and German Rigau in PATHS project; Tim Baldwin and Jey Han Lau in topic labelling. It was an honour to have a chance to work together with such great people and brilliant researchers that helped me broaden my research interests.

I would also like to thank all of my colleagues in the Natural Language Processing group for making such a nice working environment. I really appreciated helpful discussions, much needed coffee breaks and after-work drinks.

Finally, I would like to thank my family for the unreserved support and love all these years.

ABSTRACT

This thesis concerns topic models, a set of statistical methods for interpreting the contents of document collections. These models automatically learn sets of topics from words frequently co-occurring in documents. Topics learned often represent abstract thematic subjects, i.e *Sports* or *Politics*. Topics are also associated with relevant documents.

These characteristics make topic models a useful tool for organising large digital libraries. Hence, these methods have been used to develop browsing systems allowing users to navigate through and identify relevant information in document collections by providing users with sets of topics that contain relevant documents.

The main aim of this thesis is to post-process the output of topic models, making them more comprehensible and useful to humans.

First, we look at the problem of identifying incoherent topics. We show that our methods work better than previously proposed approaches. Next, we propose novel methods for efficiently identifying semantically related topics which can be used for topic recommendation. Finally, we look at the problem of alternative topic representations to topic keywords. We propose approaches that provide textual or image labels which assist in topic interpretability. We also compare different topic representations within a document browsing system.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	iv
1 Introduction	1
1.1 Contributions	4
1.2 Thesis Overview	5
1.3 Published Material	7
2 Background	8
2.1 Modelling Document Collections	9
2.2 Probabilistic Topic Models	11
2.2.1 Notation	13
2.2.2 Probabilistic Latent Semantic Analysis	14
2.2.3 Latent Dirichlet Allocation	15
2.2.4 Correlated Topic Model	18
2.3 Organising Document Collections	19
2.3.1 Limitations of Current Topic Browsers	23
2.4 Topic Coherence	23
2.5 Topic Similarity	26
2.5.1 Metrics	26
2.5.2 Applications of Topic Similarity	27
2.6 Automatic Labelling of Topics	28

2.7	Distributional Semantics	30
2.7.1	Constructing Distributional Models	31
2.8	Summary	33
3	Evaluating Topic Coherence	35
3.1	Methodology	36
3.1.1	Topic Coherence	37
3.1.2	Computing Topic Word Similarity	37
3.1.3	Constructing the Semantic Space	38
3.2	Evaluation	40
3.2.1	Data	40
3.2.2	Human Evaluation of Topic Coherence	41
3.2.3	Evaluation Metric	42
3.3	Results	44
3.4	Discussion	46
3.5	Summary	46
4	Measuring Topic Similarity	48
4.1	Methodology	49
4.1.1	Topic Word Probability Distribution Similarity	50
4.1.2	Topic Model Semantic Space	50
4.1.3	Reference Corpus Semantic Space	50
4.1.4	Training Corpus Semantic Space	51
4.1.5	Knowledge-based Methods	53
4.1.6	Feature Combination Using Support Vector Regression	54
4.2	Evaluation	54
4.2.1	Data	55
4.2.2	Generating Pairs of Topics	55
4.2.3	Human Judgements of Topic Similarity	57
4.2.4	Evaluation Metric	58
4.2.5	Baseline	58
4.3	Results	59
4.4	Summary	63

5	Automatic Labelling of Topics Using Text	64
5.1	Methodology	65
5.1.1	Generating Candidate Labels	65
5.1.2	Retrieving and Processing Text Information	66
5.1.3	Creating a Text Graph	66
5.1.4	Identifying Important Terms	67
5.1.5	Ranking Labels	67
5.2	Evaluation	68
5.2.1	Data	68
5.2.2	Evaluation Metrics	68
5.2.3	Model Parameters	69
5.3	Results and Discussion	69
5.3.1	Experimenting with the Number of Search Results	72
5.4	Summary	73
6	Automatically Labelling of Topics Using Images	74
6.1	Methodology	75
6.1.1	Selecting Candidate Images	75
6.1.2	Feature Extraction	75
6.1.3	Ranking Candidate Images	78
6.2	Evaluation	80
6.2.1	Data	80
6.2.2	Human Judgements of Image Relevance	81
6.2.3	Evaluation Metrics	82
6.2.4	Baselines	82
6.2.5	Human Performance	84
6.3	Results	84
6.4	Discussion	86
6.5	Summary	88
7	Comparing Topic Representations using an Exploratory Task	89
7.1	Methodology	90
7.1.1	Document Collection	90

7.1.2	Topic Modelling	90
7.1.3	Topic Browsing Systems	92
7.1.4	Exploratory Search Task	94
7.1.5	Subjects and Procedure	95
7.2	Results	96
7.2.1	Number of Retrieved Documents	96
7.2.2	Precision	98
7.2.3	Post-task	100
7.3	Summary	101
8	Conclusions	103
8.1	Summary of Thesis	103
8.2	Evaluation of Thesis Goals	105
8.3	Future Directions	106
	Bibliography	108

Chapter 1

INTRODUCTION

Vast amounts of information are now available on-line in digital libraries, collections and archives. Much of this information is stored in unstructured formats (such as text) and is not organised using any automated system. That is often overwhelming for users in a way that makes it very difficult to find specific information or even explore such collections.

This thesis deals with the application of a set of statistical methods, namely topic models (Blei et al., 2003; Hofmann, 1999) for analysing and organising large document collections. Topic models have been integrated into document browsing systems allowing humans to navigate through and identify relevant information on a large scale (Chaney and Blei, 2012; Gretarsson et al., 2012; Hinneburg et al., 2012).

Given a document collection, topic models learn a set of latent variables called topics. Topics are probability distributions over a vocabulary of words where frequently co-occurring words in the collection are associated with high probability. In addition, each document is represented as a distribution over topics. Table 1.1 shows a sample of topics learned from a collection of news articles represented by lists of the top-10 words with highest marginal probability in the topic.

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), and image labelling

Topic	Topic Words
1	oil, louisiana, coast, gulf, orleans, spill, state, fisherman, fishing, seafood
2	north, kim, korea, korean, jong, south, il, official, party, son
3	model, wheel, engine, system, drive, front, vehicle, rear, speed, power
4	drink, alcohol, indonesia, drinking, indonesian, four, nokia, beverage, mc-donald, caffeine
5	privacy, andrews, elli, alexander, burke, zoo, information, chung, user, regan

Table 1.1: A sample of topics generated by a topic model over a corpus of news articles. Topics are represented by top-10 most probable words.

(Feng and Lapata, 2010b).

Topic models have also proved to be a useful way to represent the content of large document collections. Visualisation interfaces (topic browsers) based on topic models have been developed in recent years (Chaney and Blei, 2012; Ganguly et al., 2013; Gretarsson et al., 2012; Hinneburg et al., 2012; Snyder et al., 2013). These systems enable users to navigate through the collection by presenting them with sets of topics. Therefore, processing and improving the output of topic models will enhance the development of document visualisation interfaces.

This thesis aims to make topic models more comprehensible and useful to humans. There are a number of challenges to overcome when applying these statistical methods to organise document collections:

- One main challenge is that topics need to present a coherent and interpretable thematic subject. There is no guarantee that each topic will be coherent. In Table 1.1, topics 1, 2 and 3 represent concrete thematic subjects. On the other hand, it is difficult to identify the underlying subjects of topics 4 and 5. Imagine how confusing these two topics would be if they are presented to users searching for specific information. Thus, providing users only with interpretable topics is an essential part of improving access to document collections organised using topic models. Finally, identifying coherent topics is required as a pre-processing step in algorithms for automatically generating topic labels (see below). That is, for a topic which

might be difficult to interpret by humans, it would be difficult to assign it an unambiguous label.

- It seems intuitively plausible that some automatically generated topics will be similar while others are dis-similar. For example, a topic about basketball (TEAM, GAME, JAMES, SEASON, PLAYER, NBA, PLAY, KNICKS, COACH, LEAGUE) is more similar to a topic about football (WORLD, CUP, TEAM, SOCCER, AFRICA, PLAYER, SOUTH, GAME, MATCH, GOAL), and golf (GOLF, WOODS, HOLE, OPEN, COURSE, SHOT, ROUND, TOUR, PLAYER, TH) than one about the global finance (FED, FINANCIAL, BANKS, FEDERAL, RESERVE, BANK, BERNANKE, RULE, CRISIS, CREDIT). Methods for automatically determining the similarity between topics have several potential applications, such as analysis of corpora to determine topics being discussed (Hall et al., 2008) or within topic browsers to decide which topics should be shown together (Chaney and Blei, 2012).
- Another challenge is to assist in the interpretation of the lists of words representing the topics by providing alternative representations. Interpreting such lists may not be straightforward, particularly since background knowledge may be required and there may not be access to documents in the source collection used to train the model. Therefore, informative labels could assist with the interpretations of topics. For example, topic 1 could be represented by a textual label, e.g. MEXICAN GULF OIL SPILL, while topics 2 and 3 could be associated with the labels NORTH KOREAN POLITICS and CARS respectively. Moreover, other media such as images could be used as labels for topics.
- Intuitively, labels, i.e. sets of keywords, textual phrases or images, represent topics in a more accessible manner than the standard keyword list approach. However, there has not been any empirical validation of this intuition. This thesis seeks to address this shortcoming. It aims to understand the impact of different topic representation modalities in finding relevant information in document collections, and also measure the level of difficulty in interpreting the same topics through different representation modalities.

1.1 Contributions

This thesis presents novel methods tackling the challenges mentioned in the previous section. Its contributions are as follows:

- Introduces novel approaches for computing topic coherence based on distributional semantics by representing words which outperform previously used methods.
- Describes a publicly available data set to evaluate topic coherence.
- Explores the problem of measuring similarity between topics by introducing topic similarity metrics including ones based on distributional semantics and knowledge-based measures. We show that the proposed metrics perform better than previously used methods.
- Introduces a publicly available data set consisting of pairs of topics together with human judgements to evaluate topic similarity.
- Proposes an unsupervised graph-based approach to associate textual labels with topics. Evaluation on a standard data set shows that the performance of the proposed graph-based method is consistently superior to previously reported methods.
- Introduces an alternative approach in which topics are represented using images. Results obtained show that the proposed approach significantly outperforms several baselines and can provide images that are useful to represent a topic.
- Describes a publicly available data set of topics and candidate image labels.
- Finally, it compares different representations for automatically-generated topics, i.e. keyword lists, textual phrases and images, within an exploratory browsing interface. Results indicate that automatically generated labels are a promising approach for representing topics within browsing interfaces.

1.2 Thesis Overview

The remainder of this thesis is organised as follows:

Chapter 2 describes the notion of topic model and introduces different types of topic models. In addition, work on organising collections of documents using topic models is introduced. Previous work on improving the output of topic models is described including methods for computing topic coherence, labelling topics and measuring topic similarity.

Chapter 3 explores methods for automatically determining the coherence of topics. It proposes a novel approach for measuring topic coherence based on the distributional hypothesis (Harris, 1954). Each topic word is represented as a bag of highly co-occurring context words that are weighted using a range of word association measures and numbers of context terms. All methods are evaluated by measuring correlation with human judgements on three different sets of topics. Results indicating that the measures proposed outperform state-of-the-art methods.

Chapter 4 explores methods for computing semantic similarity between topics. Approaches to computing topic similarity have been described in the literature but they have been restricted to using information from the word probability distribution to compare topics and have not been directly evaluated. The work in this chapter addresses these limitations by providing a systematic evaluation of a range of approaches to computing similarity between topics. It also introduces a data set consisting of pairs of topics together with human judgements of similarity to evaluate the proposed approaches. The data set has been made publicly available. Results demonstrate that the proposed methods perform better than those used previously.

Chapter 5 introduces a graph-based approach to labelling topics with textual labels. It makes use of topic keywords to form a query and retrieve relevant information from a search engine. A graph is generated from the words contained

in the search results and these are then ranked using a graph-based algorithm. Evaluation on a standard data set shows that the proposed method consistently outperforms the best performing previously reported method, which is supervised (Lau et al., 2011).

Chapter 6 presents a novel approach to labelling topics with appropriate images. The approach utilises the vast amount of pictures existing on the Web to generate a set of candidate images for each topic. Candidate images are retrieved by querying an image search engine with the top n topic terms. The most suitable image is selected by using a graph-based approach that makes use of both textual and visual information. The ranking method 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. The method is evaluated using a data set created for this study that was annotated by crowdsourcing. The data set consisting of topics and candidate images has been made publicly available. Results of the evaluation show that the proposed method significantly outperforms three baselines.

Chapter 7 compares different representations for automatically-generated topics within an exploratory browsing interface. The representations are: (1) lists of keywords, (2) textual labels, and (3) image labels. Three versions of the browsing interface were created, each using a different topic representation. An experiment is carried out in which users are asked to retrieve relevant documents using the interface. Results indicate that automatically generated labels are a promising approach for representing topics within browsing 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.

Chapter 8 summarises the main conclusions of this thesis and suggests possible avenues for future work.

1.3 Published Material

Work contributing to this thesis has been published in the following peer reviewed conferences:

- The work presented in Chapter 3 has been published in the proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) (Aletras and Stevenson, 2013a).
- The work presented in Chapter 4 has been published in the proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014) (Aletras and Stevenson, 2014a).
- The work presented in Chapter 5 has been published in the proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014) (Aletras and Stevenson, 2014b).
- The work presented in Chapter 6 has been published in the proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013) (Aletras and Stevenson, 2013b).
- The work presented in Chapter 7 has been published in the proceedings of the International Digital Libraries Conference (DL 2014) (Aletras et al., To Appear).

Chapter 2

BACKGROUND

This chapter describes how document collections can be modelled using statistical approaches. It gives a detailed description of various topic models that will be employed in the experiments presented in later chapters and systems that make use of them to organise document collections. In addition, it presents how word meaning can be modelled by the context which occurs in high dimensional vector spaces where each point represents a contextual element.

In the introduction, we informally introduced topic models and their main characteristics that make them appropriate for organising large collections of text documents. We also indicated the main problems of applying such statistical methods, namely: (1) identifying topics that are difficult to interpret by humans, (2) summarising the main thematic subject of topic keywords in a condensed form (i.e. textual label) and (3) identifying semantically similar topics. We argued that tackling these problems can improve the output of topic models making them easier to use.

The chapter begins by presenting methods to modelling document collections in Section 2.1. Section 2.2 introduces the mathematical formulation of topic models. Next, Section 2.3 describes systems that make use of topic models to organise and visualise document collections. Also, we discuss previous approaches on measuring topic coherence (Section 2.4), summarising topics using labels (Section 2.6) and identifying semantically similar topics (Section 2.5) in the following three sections. Section 2.7 introduces vector space models of word meaning. Finally, a

summary (Section 2.8) highlights the key points of this chapter.

2.1 Modelling Document Collections

A fundamental problem in NLP is finding ways to represent large amounts of text in a compact way. One of the most common semantic representations of a document collection is the Vector Space Model (VSM) (Salton et al., 1975). In a VSM, also referred as a *semantic space* or *bag-of-words* representation, documents and terms are represented as points in a Euclidean space. The term-document matrix (Salton and McGill, 1983; Salton et al., 1975; Turney and Pantel, 2010) C is a type of a semantic space. It is a $W \times D$ matrix that contains information about the occurrence (frequency) of W terms in D documents. Elements, c_{ij} in C represent the i^{th} word in the j^{th} document and are usually weighted by the raw frequency of terms in documents or using the tf-idf (term frequency - inverted document frequency) weighting (Salton and McGill, 1983).

A well-known problem with the term-document VSMs is the high dimensionality caused by the large number of unique terms¹ that can exist in a corpus. High dimensional spaces are often also sparse, i.e. many documents contain only a few unique terms or some terms appear only in a few documents. This makes it difficult, for example, to accurately compute the similarity between two documents or terms. In addition, VSMs do not cope with polysemy since ambiguous words are represented as a single vector in the semantic space. For example, the term *python* may occur in documents about either *snakes* or *programming languages*, however there is only one instance of that vector which contains the co-occurrence of that word with all the documents in the collection.

Statistical methods such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer and Dumais, 1997) have been used to reduce the dimensionality of semantic spaces. LSA applies Singular Value Decomposition (SVD) to the matrix C . This is a form of factor analysis where C is decomposed into

¹In the rest of the thesis, we will refer to unique terms or word types.

three other matrices:

$$C = U\Sigma V^T \quad (2.1)$$

where U is a $W \times W$ matrix of word vectors where its columns are eigenvectors of CC^T , Σ is a diagonal $W \times D$ matrix containing the singular values and V is a $D \times D$ matrix of document vectors where its columns are eigenvectors of C^TC . The multiplication of the three component matrices results in the original matrix, C . Any matrix can be decomposed perfectly if the number of singular values is no smaller than the smallest dimension of C . When fewer singular values are used then the matrix product is an approximation of the original matrix. LSA reduces the dimensionality of the SVD by deleting coefficients in the diagonal matrix Σ starting with the smallest. This method achieves high compression of the original vector space which is useful in large document collections. The approximation of matrix C retaining the K largest singular values is then given by:

$$C \approx U_K \Sigma_K V_K^T \quad (2.2)$$

where U_K is a $W \times K$ matrix of word vectors, Σ_K is a $K \times K$ diagonal matrix with singular values and V_K is a $K \times D$ matrix of document vectors.

LSA can be considered as a type of topic model (Stevens et al., 2012) since it learns a set of topics, T_K , by multiplying the word vectors, U_K , and the diagonal matrix Σ_K :

$$T_K = U_K \Sigma_K \quad (2.3)$$

In addition, matrix V_K^T corresponds to the assignment of topics in each document.

LSA has been shown to capture linguistic notions such as synonymy and polysemy (Landauer and Dumais, 1997). However, Stevens et al. (2012) showed that topics learned by LSA are not easily interpretable. Vectors in T are linear combinations of the term-document frequencies and consist of positive and negative values. Therefore, it is not possible to identify important terms that characterise the main thematic subject of the topics. On the other hand, topics learned by an alternative type of topic model, described in the following section, are more

descriptive and often represent a coherent subject.

2.2 Probabilistic Topic Models

Probabilistic topic models (Blei et al., 2003; Hofmann, 1999), often referred to simply as topic models, are generative models that learn a set of latent variables called topics. The main assumption of topic models is that documents are generated by a mixture of topics while topics are probability distributions over words. The input of a topic model is a set of documents and its output is a set of topics together with topic assignments to documents.

Figure 2.1 shows an overview of a topic modelling pipeline including input and output. In the input, each document is represented as a bag-of-words. Each document is often tokenised (split up) into words and normalised (converted to lower case) while word order is ignored. The only information relevant to the model is the number of times a word appears in each document. The input section in Figure 2.1 shows documents represented as bag-of-words. Document 1 contains information about Python (programming language) while document 2 contains information about pythons (snakes).

The output of a topic model is a set of topics and a set of topic assignments for each document. Each topic is a probability distribution over all the unique words in the collection. Topics are often represented by the words with the highest probability in the topic (see the output in Figure 2.1). In the rest of the thesis, the term *topic words* refers to the set of n words with highest probability in a given topic. Words assigned high probability in some topics frequently appear together in documents and are likely to represent a coherent subject or theme.

A document is represented as a probability distribution over topics with only a few topics assigned with high probability. In Figure 2.1 showing the output, topic 1 which is about snakes is assigned with high probability to document 2. On the other hand, topic 2 which represents information about programming is assigned to document 1.

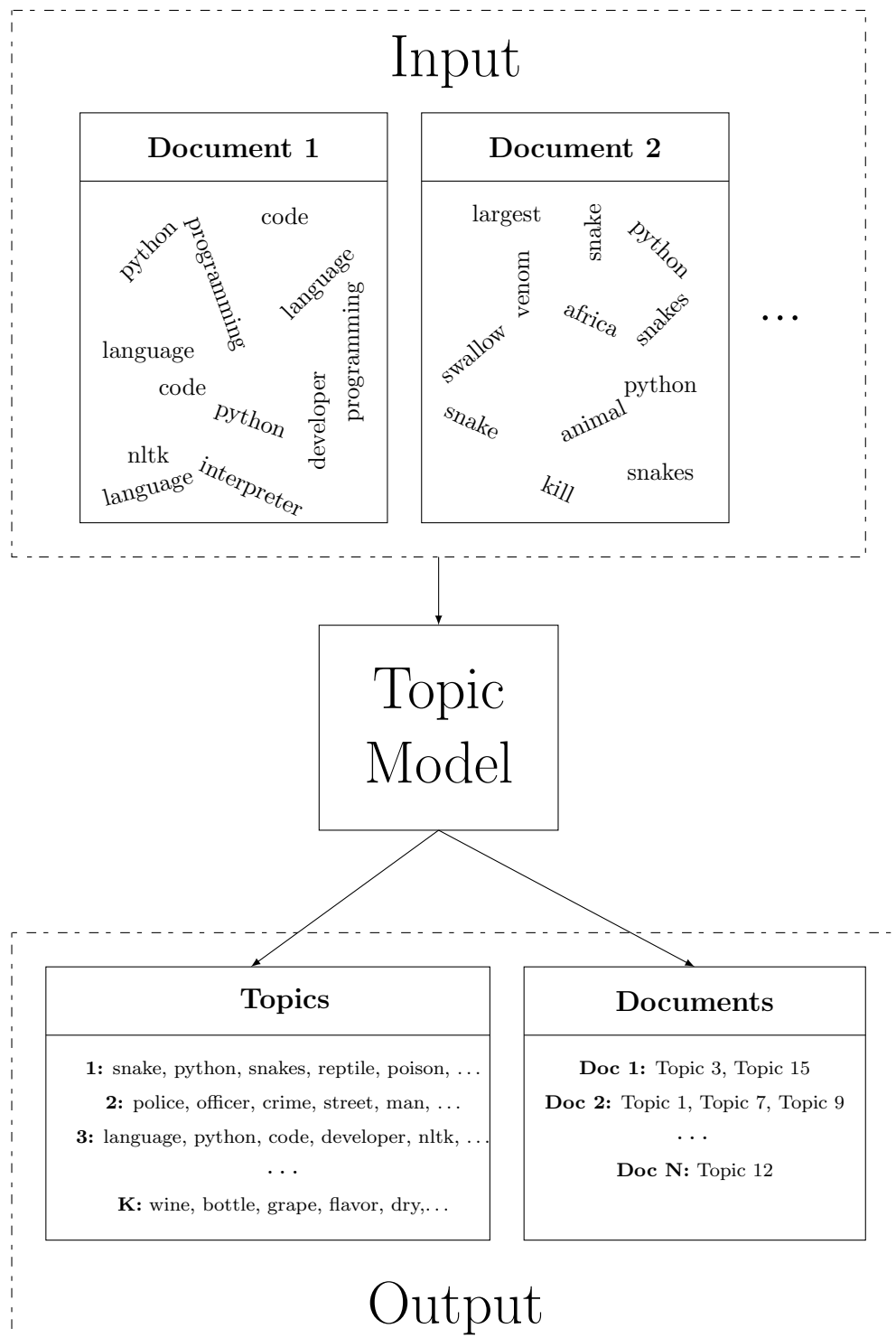


Figure 2.1: Input and Output of a Topic Model.

Topic models have appealing characteristics for organising document collections. They can be used for clustering documents under topics enhancing browsing and improving information access. In addition, topic models soft cluster terms into topics dealing with polysemy (e.g. topics 1 and 2 represent different senses of the word *python*) and therefore, users can retrieve different sets of documents relevant to a given search term (e.g. *python*).

2.2.1 Notation

We introduce topic modelling notation following a similar approach as Blei et al. (2003):

- A *word* or term represents a unique word type of a fixed length vocabulary indexed by $\{1, \dots, W\}$. Each word is represented as unit-basis vector of length W that has a single element equal to one and all other elements equal to zero. The k -th word in the vocabulary is represented by a vector w such that $w^k = 1$ and $w^i = 0$ for $i \neq k$.
- A *document* of N words is represented as a sequence by $\mathbf{d} = (w_1, w_2, \dots, w_N)$, where w_i is the i -th word in the sequence. Note that this is also a bag-of-words representation since the word sequence does not need to match the original word order of the document.
- A *corpus* is a collection of D documents $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_D\}$.

Lets consider an example vocabulary, $v = \{\text{be,not,or,to}\}$ with indices $\{1, 2, 3, 4\}$. The word *be* is represented as $w^3 = \langle 1, 0, 0, 0 \rangle$. The document $d = \text{“to be or not to be”}$ will be represented as $\mathbf{w}_d = (w_1^4, w_2^1, w_3^3, w_4^2, w_5^4, w_6^1)$, using the notation described above. Note that any permutations of the word order in \mathbf{w}_d do not have any effect and result in equal representations of the document.

We will also use $P(z|\mathbf{d})$ to denote a document’s distribution over topics, $P(w|z)$ the probability distribution over words w given the topic z and $P(w|\mathbf{d})$ the distribution over words within document D . For a corpus with D documents and W words, a topic model learns a relation between words and topics, T , and a relation between topics and documents as:

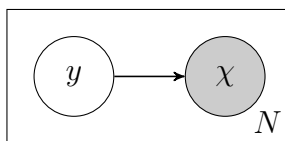


Figure 2.2: Example of plate notation.

- a $T \times W$ matrix, Φ with elements ϕ_{ij} denoting the probability $P(w_i|z = j)$, and
- a $D \times T$ matrix, Θ , with elements θ_{ij} denoting the probability $P(z = j|\mathbf{d}_i)$.

Probabilistic generative models can conveniently be represented using plate notation. In these graphical models, observed variables are indicated using shaded nodes while latent variables are denoted by unshaded nodes. Arrows between nodes indicate conditional dependency while plates (boxes) surrounding nodes indicate repetitions of sampling steps. The number in the bottom right corner of the plate indicates the number of samples (repetitions). Figure 2.2 shows an example of plate notation of a simple graphical model where χ is an observed variable and y is a latent variable.

2.2.2 Probabilistic Latent Semantic Analysis

Hofmann (1999) proposed the probabilistic Latent Semantic Analysis (pLSA) as an alternative to LSA. The pLSA models each word in a document as a sample of a mixture model. The mixture model is a set of *topics* in the form of multinomial random variables.

Given T topics, the aim is to find the probability distribution of words in a topic and the probability distribution of topics in a document. Words are the observed variables while topics are the latent variables. The generative process, illustrated in Figure 2.3, is as follows:

1. For each document $\mathbf{d} \in \mathcal{D}$ with probability $P(\theta_d)$,
 - (a) select a latent topic z with probability $P(z|\mathbf{d})$,
 - (a) generate a word w with probability $P(w|z)$.

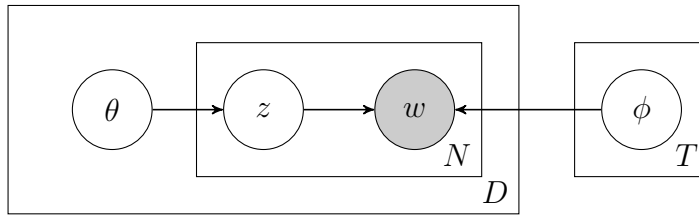


Figure 2.3: Graphical model representation of pLSA.

The above process is defined by the following expression as a joint probability between a word and a document:

$$P(\theta_d, w) = P(\theta_d)P(w|\theta_d), \quad (2.4)$$

$$P(w|\theta_d) = \sum_{z \in Z} P(w|z)P(z|\theta_d) \quad (2.5)$$

$$= P(\theta_d) \sum_{z \in Z} P(w|z)P(z|\theta_d). \quad (2.6)$$

The pLSA model satisfies the main assumption of topic models, namely that a document consists of multiple topics. The probability $P(z|\theta_d)$ contains the weight of each topic $z \in Z$ given a document \mathbf{d} . However, representing each document as a list of topic weights without a generative probabilistic model for them leads to two main problems (Blei et al., 2003): (1) the number of parameters grows linearly with the number of documents in the corpus causing overfitting problems, and (2) it is not possible to assign topic probabilities to any unseen documents, i.e. not in the training corpus.

2.2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is an extension of pLSA which introduces symmetric Dirichlet priors on the distribution over topics for a particular document, θ , and the distribution over words for a particular topic, ϕ . That addresses the problems with pLSA mentioned above by treating the topic weights in each document as a hidden random variable of size T , where T is the number of topics. A graphical representation of LDA is shown in Figure 2.4.

The generative process for the topics is as follows:

1. For each topic
 - (a) Choose a distribution over words $\phi \sim Dir(\beta)$
2. Choose N

A document \mathbf{d} in a corpus \mathcal{D} is represented by latent topics using the following generative process:
3. Choose $\theta \sim Dir(\alpha)$
4. For each of the N words w_n
 - (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
 - (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

where:

- N is the number of words in a document.
- z_n is the n topic for the word w_n .
- θ is the topic distribution for a document.
- α is the parameter of the Dirichlet prior on the per-document topic distributions.
- β is the parameter of the Dirichlet prior on the per-topic word distribution.

The joint probability of the corpus \mathcal{D} given the hyperparameters α and β is given by the following equation:

$$P(\mathcal{D}|\alpha, \beta) = \prod_{t=1}^T \prod_{d=1}^D \prod_{n=1}^N P(\phi_t|\beta)P(\theta_d|\alpha)P(z_{dn}|\theta_d)P(w_{dn}|\phi_{z_{dn}}) \quad (2.7)$$

The main variables need to be estimated in the model are the per-topic word distributions ϕ and the per-document topic distributions θ . Inferring direct estimates from Equation 2.7 is intractable.

Hofmann (1999) used the expectation-maximisation (EM) algorithm to estimate ϕ and θ directly. However the EM algorithm may get stuck in local

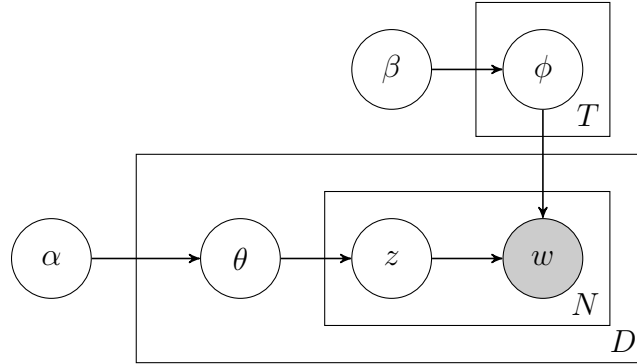


Figure 2.4: Graphical model representation of LDA.

maxima. Approximation methods have been used to avoid this problem such as Bayesian variational inference (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004).

Gibbs sampling is a specific type of Markov Chain Monte Carlo model (MCMC) for obtaining sample values from complex multivariate probability distributions. It starts by assigning every word w with a random topic $t \in \{1, \dots, T\}$ for every document in corpus \mathcal{D} . Then, for each word, it estimates the probability of assigning the current word in each topic, given the topic assignments of all of the other words. Griffiths and Steyvers (2004) calculated this probability by:

$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^W C_{w j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^T C_{d_i t}^{DT} + T\alpha} \quad (2.8)$$

where $z_i = j$ represents the topic assignment of word w_i in topic j , \mathbf{z}_{-i} are the topic assignments of all of the other words, and “.” represents all the other information from word, documents and hyperparameters α and β . In addition, the element $C_{w j}^{VT}$ of the matrix C^{VT} , $V \times T$, contains the number of times word w assigned to topic j . The matrix C^{DT} , $D \times T$, contains the counts of each topic is assigned to words of each document ; $C_{d j}^{DT}$ represents the number of times topic j is assigned to words in document \mathbf{d} .

The Gibbs sampling algorithm estimates the probability of each topic for every word. The elements of matrices Φ and Θ , containing the per-topic word

distributions and the per-document topic distributions, can be obtained by:

$$\phi_{ij} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^W C_{kj}^{WT} + W\beta} \quad \theta_{jd} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^T C_{dk}^{DT} + T\alpha} \quad (2.9)$$

where ϕ_{ij} is the probability of word w^i in topic j and θ_{jd} is the probability of topic j in document \mathbf{d}_d .

2.2.4 Correlated Topic Model

LDA assumes that topics are independent and does not attempt to capture correlations between them. The Correlated Topic Model (CTM) (Blei and Lafferty, 2006) overcomes this limitation by capturing topic correlation via the logistic normal distribution. Let $\{\mu, \Sigma\}$ be a K -dimensional mean and covariance matrix, and let topics $\phi_{1:T}$ be T multinomials over a vocabulary of size W . An N -word document is generated by the following process shown in Figure 2.5:

1. Choose $\theta | \{\mu, \Sigma\} \sim \mathcal{N}(\mu, \Sigma)$
2. For each of the N words w_n
 - (a) Choose a topic $Z_n | \eta$ from $\text{Mult}(f(\theta))$
 - (b) Choose word $w_n | \{z_n, \phi_{1:T}\}$ from $\text{Mult}(\phi_{z_n})$

In CTM, the distribution over topics for a document is drawn from a logistic normal distribution. The covariance matrix for parametrising the logistic normal distribution can be used to identify correlations between topics and form a topic graph in which each node represents a topic and each edge denotes the correlation between them.

Apart from the fact that CTM identifies correlations between topics which LDA does not capture, it has also been shown to achieve higher predictive likelihood than LDA (Blei and Lafferty, 2006). On the other hand, training LDA models is less computationally intensive. In addition, Chang et al. showed that LDA produces topics that are more coherent and easy to interpret by humans than those produced by CTM.

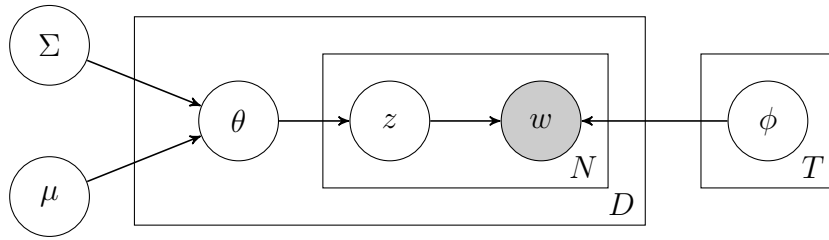


Figure 2.5: Graphical model representation of CTM.

2.3 Organising Document Collections

So far, we have shown that topic models decompose a document collection into a per-document topic matrix, Θ , and a per-topic word matrix, Φ . Θ can be considered as a soft-clustering of documents under topics. Each column vector in it contains the probability of a particular topic in each document. Therefore, topic models can be naturally used to automatically organise and visualise document collections as sets of topics. Users can access documents by navigating through topics. The remainder of this section describes various document collection visualisation and analysis tools based on topic modelling.

Topic Browser (Gardner et al., 2010) is a topic-based browsing system which provides visualisation of a document collection by presenting topics and associated documents. Apart from the information from the topic model, Topic Browser provides a broad range of metrics such as the number of words labelled with a particular topic or the spread of topics across documents. The system also computes topic coherence and similarity, as well as similarity between documents.

Newman et al. (2010a) proposed a different approach to the visualisation of a topic model using maps. The topic mapping tool, Topic Maps, takes as an input the matrix Θ which contains the per-document topic distributions and applies further dimensionality reduction techniques to project it onto a two-dimensional space. Each topic is mapped to a colour while documents are represented by the colour of the most probable topic given the document. In that way, similar documents appear close together on the projected space. Users can browse documents accessing the map or by clicking on text links on the right side of the map.

TIARA (Text Insight via Automated Responsive Analytics) (Wei et al., 2010)

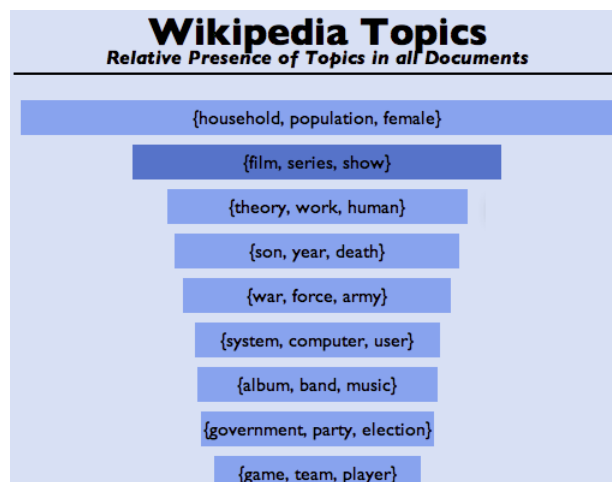
is a system designed to visualise the topics in a document collection. Users can interactively view, explore, and analyse text through plots of topic evolution over time. Topics are represented by sets of keywords and users can access text of specific documents under each topic.

Chaney and Blei (2012) presented TMVE (Topic Model Visualisation Engine), a document collection visualisation tool based on LDA. Given a set of documents and a trained LDA model, the tool generates an interface with the following main components: a main page showing a list of topics, topic pages and document pages. The main page contains the list of generated topics represented by a set of keywords. Each topic page is associated with 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. Figure 2.6 shows parts of TMVE's interface.

Another web-based system providing analysis and visualisation capabilities using topic models is TopicNets (Gretarsson et al., 2012). TopicNets represents documents and topics as nodes of different types in an interactive graph. Edges connecting document and topic nodes are weighted by the probability of the topic given the document, θ_{td} while edges connecting topics are weighted by the similarity of their word distributions.

Hinneburg et al. (2012) developed a system called TopicExplorer. It allows exploration of a collection of Wikipedia documents and offers search and visualisation capabilities. Topics are presented by sets of keywords together with image thumbnails extracted from documents that have high probability given the topic. Topics are presented in a linear order, mapped to colour scale with similar topics placed close. Documents are also associated with colours based on their per-topic distribution. In addition, users can retrieve documents and create a shortlist. Given the average topic mixture of the shortlist, the system provides recommendations of potentially relevant documents. Figure 2.7 shows the TopicExplorer interface showing topics and documents retrieved for the keyword *football*.

Chuang et al. (2012) developed the Stanford Dissertation Browser, a visual analysis tool for exploring PhD theses from 75 departments in Stanford University. It follows a similar approach to topic maps (Newman et al., 2010a)

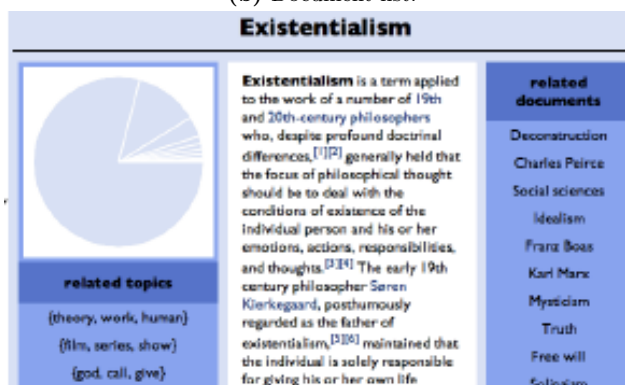


(a) Topic list.

{school, student, university}

words	related documents	related topics
school	College	{day, year, event}
student	High school	{rate, high, increase}
university	Education in the United States	{service, military, aircraft}
college	Columbia University	{work, book, publish}
high	Juris Doctor	{group, member, jewish}
program	Ohio State University	{car, race, vehicle}
serve	Georgetown University	{village, small, smallsup}
grade	University of California, Berkeley	{city, large, area}
campus	Rutgers University	{land, century, early}
education	University of Florida	{area, part, region}
member	University of Chicago	{water, park, boat}
year	University of California, Los Angeles	{style, bgcolor, rowspan}

(b) Document list.



(c) Document page.

Figure 2.6: Topic Model Visualisation Engine (Chaney and Blei, 2012).

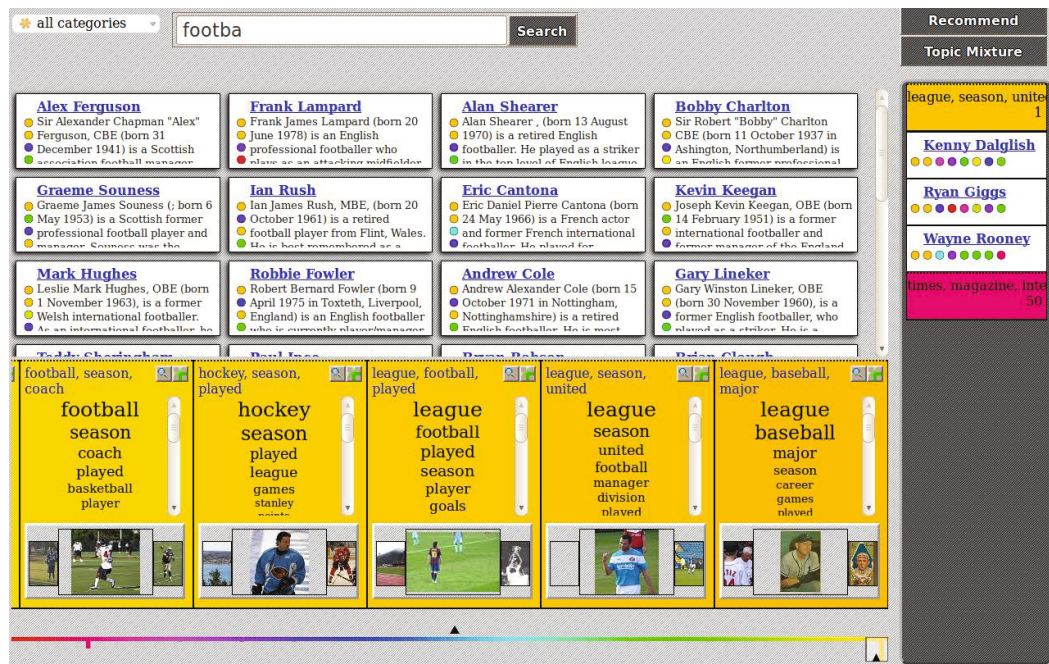


Figure 2.7: TopicExplorer (Hinneburg et al., 2012)

projecting departmental and thesis relationships into a two-dimensional space. It supports multiple views by measuring similarity of topic distributions of these or departments given an LDA model.

TopicVis (Ganguly et al., 2013) is a system for topic-based navigation. It retrieves a set of relevant documents given a user query and applies LDA over that set. The obtained per-topic word distribution ϕ is used to create a pie chart of the topic distribution of the retrieved documents. Each topic is represented by the 10 most probable words given the topic. Furthermore, each document in the retrieved list is associated with a bar chart showing the distribution of topics given the document.

Snyder et al. (2013) presented MetaToMATo, a web-based system for topic browsing. It combines automatically generated topics from a topic model with a metadata-based system. The system offers topic management, document filtering by topic and summarised views which contain metadata together with topic graphs.

More recently, Sievert and Shirley (2014) presented LDAvis, a web-based in-

terface for topic visualisation using R² and D3³ while Smith et al. (2014b) described a relationship-enriched visualisation system to help users explore topic models through word and topic correlations. In addition, Smith et al. (2014a) presented, Hierarchy, a tool for visualising hierarchical topic models.

2.3.1 Limitations of Current Topic Browsers

A common characteristic of the majority of the systems described above is that they do not remove incoherent topics (the only exception is the Topic Browser (Gardner et al., 2010)). Therefore, it is likely that users are provided with meaningless sets of words making navigation difficult. In addition, none of the systems provide alternative representations of topics. Topics are presented by sets of keywords. Finally, some of the systems identify similar topics, however using naive methods such as computing the similarity of the per-topic word probability distributions without empirically evaluating the similarity metrics. The following sections describe previous work on dealing with these problems.

2.4 Topic Coherence

There is no guarantee that all of the topics generated by a topic model will represent a coherent subject. It is likely that some of them will be meaningless or “junk”. The systems described in the previous section that do not consider topic coherence may provide users with junk topics, which could make navigation an unpleasant experience. Let’s consider two topics, $t_1 = \{\text{AFRICA, SOUTH, MOBUTU, AFRICAN, ZAIRE, KABILA, PRESIDENT, MANDELA, KINSHASA, CONGO}\}$ and $t_2 = \{\text{THINK, LIKE, ONE, GOING, GET, GOOD, RE, TIME, VE, BACK}\}$. Topic t_1 can easily be interpreted as “Africa” or “African Affairs”. Conversely, t_2 consists of words which are difficult to interpret as a coherent unit. Methods for automatically identifying coherent and incoherent topics can dramatically improve the output of a topic model in a visualisation interface.

Automatically computing topic coherence has been also proved a standard

²<http://www.r-project.org/>

³<http://d3js.org/>

way of evaluating topic models (Chang et al.). In early research on topic models evaluation, extrinsic methods were used and the model's performance measured by applying it to a specific task. For example, topic models have been evaluated by measuring their accuracy for information retrieval (Wei and Croft, 2006). Statistical methods have also been applied to measure the predictive likelihood of a topic model in held-out documents by computing their perplexity (Wallach et al., 2009).

However, these approaches do not provide any information about how interpretable the topics are to humans. AlSumait et al. (2009) describe the first attempt to automatically evaluate topic coherence. Three criteria are applied to identify junk or insignificant topics. The criteria are in the form of probability distributions over the highest probability words. For example, topics in which the probability mass is distributed approximately equally across all words are considered likely to be difficult to interpret. Chang et al. showed that humans find topics generated by models with high predictive likelihood, i.e. CTM, to be less coherent than topics generated from others with lower predictive likelihood, i.e. LDA. Following Chang's findings, recent work on evaluation of topic models has focused on automatically measuring the coherence of generated topics and, incorporating such methods into the topic models.

Newman et al. (2010b) proposed a method for automatically computing topic coherence which has been shown to be highly correlated with human evaluation. It is assumed that a topic is coherent if all or the most of its top-10 words are related. Results showed that word relatedness is better predicted using the distribution-based Pointwise Mutual Information (PMI) (Church and Hanks, 1989). PMI was computed using Wikipedia as an external reference corpus and proved to work better than knowledge-based measures. Mimno et al. (2011) showed that available co-document frequency of words in the training corpus can be used to measure semantic coherence. Topic coherence is defined as the sum of the log ratio between co-document frequency and the document frequency for the N most probable words in a topic. The intuition behind this metric is that the co-occurrence of words within documents in the corpus can indicate semantic relatedness. On the other hand, Musat et al. (2011) associated words in a topic with WordNet concepts thereby creating topical subtrees. The method relies on Word-

Net’s hierarchical structure to find a common concept that best describes as many words as possible. It is assumed that the higher the coverage and specificity of a topical subtree, the more semantically coherent the topic. Experimental results showed high agreement with humans in the word intrusion task where humans were presented with lists of keywords representing topics and asked to find the word which is less relevant to the others (intruder). In contrast, Newman et al. (2010b) concluded that WordNet is not useful for evaluation of topic models. In a similar fashion, topic words can be mapped onto the Wikipedia page-links graph and graph-centric features can be extracted from the sub-graph defined by the articles corresponding to topic words (Chan and Akoglu, 2013). Topic coherence is computed using graph-based features in a supervised approach.

The metrics proposed above can be used to compute the coherence of complete topic models, i.e. coherence of the entire set of topics generated by a topic model. Stevens et al. (2012) evaluated complete topic models by computing the coherence of each individual generated topic. They integrated the methods proposed by Newman et al. (2010b) and Mimno et al. (2011) into aggregated metrics. Results showed that LDA generates more coherent topics than LSA.

Researchers have also made efforts to improve existing topic models or develop new ones that generate more coherent topics. Andrzejewski et al. (2009) proposed a method for generating coherent topics which use a mixture of Dirichlet distributions to incorporate domain knowledge. Their approach prefers words that have similar probability (high or low) within all topics and rejects words that have different probabilities across topics. Other researchers incorporated topic coherence metrics into topic models to produce more interpretable topics (Chen et al., 2013; Mimno et al., 2011; Newman et al., 2011).

Recent work by Ramirez et al. (2012) analyses and evaluates the semantic coherence of the results obtained by topic models rather than the semantic coherence of the inferred topics. Each topic model is treated as a partition of document-topic associations and evaluated using metrics for cluster comparison.

Identifying coherent topics is an important stage in post-processing the output of topic models. Its importance are twofold: (1) provides topic model evaluation and (2) improves visualisation interfaces by filtering out topics that cannot be interpreted by humans.

2.5 Topic Similarity

It seems intuitively plausible that some automatically generated topics will be similar while others are dis-similar. Topic similarity can be used for topic recommendation. For example, users accessing a topic about cinema may be interested in a topic about television series. In the same fashion, similar documents can be identified and recommended. On the other hand, topic similarity can be used for comparing large corpora to automatically identify semantic overlaps.

2.5.1 Metrics

Per-topic word distributions in Φ can be naturally used to measure topic similarity. Previous work on measuring similarity between topics relied on approaches that compare the topics' word probability distributions. A common function to measure the difference of two probability distributions, P and Q , is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951):

$$D_{KL} = \sum_i P(i) \log_2 \left(\frac{P(i)}{Q(i)} \right) \quad (2.10)$$

The KL-divergence is asymmetric. Jensen-Shannon Divergence (JSD) is a symmetric measure that computes the distance between two probability distributions as the distance from each one to the mean distribution. JSD is defined as follows:

$$\text{JSD}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \quad (2.11)$$

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log_2 \left(\frac{P(i)}{Q(i)} \right) \quad (2.12)$$

where $M = \frac{P+Q}{2}$ is the mean of distributions P and Q .

Since JSD measures distance rather than similarity, the similarity between

two topics can be defined as follows:

$$\text{Sim}_{\text{JSD}}(\phi_i, \phi_j) = \begin{cases} 1 - \text{JSD}(\phi_i \parallel \phi_j) & \text{if } \text{JSD}(\phi_i \parallel \phi_j) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.13)$$

where ϕ_i and ϕ_j is the word probability distribution of the topics i, j respectively where $\phi_i, \phi_j \in \Phi$.

2.5.2 Applications of Topic Similarity

Li and McCallum (2006) generated a set of pairs of similar topics consisting of one LDA topic and one topic generated using the Pachinko Allocation Model. The aim of the experiment was to determine which of the two models generates more semantically coherent topics by asking human subjects to select the most coherent topic from a given pair. Similarity between topics is measured as the KL-divergence between their word distributions. Wang et al. (2009) proposed a model that relied on random initialisation and also used KL-divergence to compare the topics that were generated by different runs to determine how similar they were. Newman et al. (2009) introduced distributed topic models which are trained on multiple processors. The topics created on each processor are merged by computing their similarity. Topic similarity is defined as the symmetric KL-divergence between two topics. Note that the symmetric KL-divergence is slightly different to JSD and is computed as $\frac{1}{2}(D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P))$. The same approach has been followed by Gretarsson et al. (2012) to create a topic-similarity layout for a text visual analysis tool called *TopicNets* (see Section 2.3).

He et al. (2009) computed similarity between two topics as the cosine of the angle of their word distributions. This measure of topic similarity was used to monitor the evolution of topics in scientific literature. Ramage et al. (2009) used the same measure to compute the similarity of topics learned by supervised and unsupervised models (Labelled-LDA and standard LDA) generated over the same corpus.

Chaney and Blei (2012) computed similarity between topic distributions using a formula similar to the average Log Odds Ratio and applied it in their topic browser.

The similarity metrics used in the work described above have not been directly evaluated. Kim and Oh (2011) measured similarity between topics by applying LDA to news articles in different time frames. The main aim of that work was to identify topic chains which monitor the time a topic is in the news. They applied six measures of similarity between the word distributions obtained from pairs of topics: Cosine Similarity, Jaccard's Coefficient, Kendall's τ coefficient, Discount Cumulative Gain, KL-Divergence and Jensen Shannon Divergence (JSD). Results shown that JSD provides the best estimates of topic similarity. However, their evaluation was indirect and involved substituting similar topics from one model to another then evaluating the new model.

The metrics described above have not been formally evaluated against human judgements. Chapter 4 compares a broad range of metrics for estimating topic similarity against human judgements.

2.6 Automatic Labelling of Topics

It is often useful to present topics in a summarised form. One common approach is to automatically generate textual labels. These labels summarise topics' main thematic subject and provide a more convenient representation compared to a list of keywords. 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 convenient presentation of research results (Mei and Zhai, 2005; Teh et al., 2006). Table 2.1 shows examples of topics represented by lists of keywords and textual labels.

The first attempt at automatically assigning labels to topics is described by Mei et al. (2007). In their approach, a set of candidate labels is extracted from a reference collection using noun chunks and statistically important bigrams. A relevance scoring function is defined which minimises the distance 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 to labelling topics that relies on two hierarchical knowledge resources labelled by humans, the Google Directory

Label	Topic (Top-10 Terms)
TV and Media	show, television, tv, news, network, medium, fox, cable, channel, series
European Union/Euro-zone	european, euro, europe, country, germany, union, ireland, french, france, government
Music	song, music, band, album, rock, pop, record, singer, sound, guitar
UK Politics	party, britain, minister, british, prime, government, london, conservative, cameron, liberal
Fashion	look, dress, wear, shirt, hair, fashion, wearing, style, shoes, black

Table 2.1: A sample of topics generated by a topic model over a corpus of news articles together with appropriate labels. Topics are represented by top-10 most probable words.

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. The label is selected by computing the similarity between each word and all others in the topic. Several sources of information are used to identify the best label including Pointwise Mutual Information scores, WordNet hypernymy relations and distributional similarity. These features are combined in a reranking model 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. A set of candidate labels is generated in four phases. First, 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 (Grieser et al., 2011) 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. 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. Results show that incorporating information about the relations between topics improves label quality.

Hulpus et al. (2013) make use of the structured data in DBpedia⁴ to label topics. Their approach maps topic words to DBpedia concepts and identifies the best ones by applying graph centrality measures assuming that words co-occurring in text likely refer to concepts that are closer in the DBpedia graph.

More recently, Cano Basave et al. (2014) presented a method for labelling LDA topics trained on social media streams, i.e Twitter, using summarisation techniques. Their method generates labels which exist in the Twitter stream rather than relying on external knowledge sources.

Topic labelling methods summarise topics into a condensed form. Usually, labels are unique words, bigrams, keyphrases or Wikipedia article titles. Chapter 5 presents an approach to labelling topics which achieves state-of-the-art performance. Previous work on the task has been entirely focused on generating labels from only one medium. Chapter 6 proposes an alternative representation of topics using images while Chapter 7 compares the effectiveness of various topic representations.

2.7 Distributional Semantics

Intuitively, one can often guess the meaning of a word from its context. This has been formally defined as the distributional hypothesis which states that words

⁴<http://dbpedia.org>

with similar meanings tend to occur in similar context (Firth, 1957; Harris, 1954). Distributional models represent words as vectors of co-occurrence counts in contexts, i.e. words, sentences, paragraphs, documents, syntactic patterns or visual attributes (Clark, 2012; Erk, 2012; Turney and Pantel, 2010). This is also known as a term-context matrix.

Section 2.1 presented the term-document matrix, C , for modelling document collections. Matrix C can also be seen as special case of a word-context matrix, i.e. by looking at row vectors, where each context feature is an entire document. The term-context matrix is an extension of the term-document matrix to represent words in high dimensional vector spaces.

Erk (2012) presents a wide range of NLP tasks where distributional semantics have been applied. Clark (2012) and Erk (2012) describe compositionality where the meaning of a phrase can be represented as a vector by combining its constituent word vectors. In the rest of the thesis, we make use of distributional semantics to represent topic words in various semantic spaces. For example, we compute topic coherence (see Chapter 3) and similarity (see Chapter 4) by measuring distributional similarity of topic words.

2.7.1 Constructing Distributional Models

Constructing a word-context matrix requires the definition of various parameters. The main parameters are a basis, a weighting function, a similarity metric and a transformation (Lowe, 2001).

The basis of a term-context matrix contains the vector elements, also known as context features, which represent the context where the co-occurrences are observed. The basis often consists of a set of words. When words are used as context features, a context window of specific length around the target word is defined. This can be a fixed number of words on either sides of the target word or an entire sentence, paragraph or document. Apart from using words, more recent work attempted to represent words using visual features extracted from images occurring naturally with text (Bruni et al., 2011; Feng and Lapata, 2010a; Kiela et al., 2014; Lazaridou et al., 2014). Figure 2.8 shows a small toy corpus, a context vocabulary and the resulting vector space of the words *tablet*, *laptop*,

movie and *film*.

The weighting function transforms the raw co-occurrence counts to word association weights to reduce noise. Weighting is used to assign less weight to words co-occurring by chance or less informative very frequent words, i.e. stop words. A popular word association measure which will be used extensively in the rest of the thesis is the Pointwise Mutual Information (PMI) (Church and Hanks, 1989). It computes the variation between the probability of the co-occurrence of two words given their joint distribution, $p(w_i, w_j)$, and their individual distributions, $p(w_i)$ and $p(w_j)$, assuming independence. PMI is computed as follows:

$$\text{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (2.14)$$

$$= \log_2 \frac{c(w_i, w_j) \times N}{c(w_i)c(w_j)} \quad (2.15)$$

where $c(w_i, w_j)$ is the number of co-occurrences of w_i and w_j , $c(w_x)$ is the frequency of w_x in C and N is the size of C . In addition, vectors are also weighted using **NPMI** (Normalised PMI). This is an extension of PMI that has been used for collocation extraction (Bouma, 2009) and is computed as follows:

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log_2(p(w_i, w_j))} \quad (2.16)$$

A common practice in the distributional semantics literature is to convert to zero all the negative PMI and NPMI values (Clark, 2012; Erk, 2012; Turney and Pantel, 2010).

The transformation usually transforms the whole vector space, i.e. by applying dimensionality reduction techniques. For example, SVD (Section 2.1) or non-negative matrix factorisation (Clark, 2012; Turney and Pantel, 2010).

Finally, semantic similarity between words is computed using a similarity metric between a pair of vectors. Popular metrics include the cosine of the angle of the word vectors, Dice and Jaccard coefficients (Manning et al., 2008).

A tablet computer is a mobile personal computer that is primarily operated by touching the screen .

A laptop is a portable personal computer with a clamshell form factor, suitable for mobile use .

The movie streaming option will work on tablets allowing users watching the latest dramas and comedies .

Comedy is a genre of film in which the main emphasis is on humour .

George's favourites are dramas and action films .

According to critics, Snatch is one of the best action - comedy movies of 00s .

Context vocabulary: *computer, screen, personal, drama, comedy, action*

	computer	screen	personal	drama	comedy	action
tablet	2	1	1	1	1	0
laptop	1	1	1	0	0	0
movie	0	0	0	1	1	1
film	0	0	0	1	1	1

Figure 2.8: A small toy corpus of six sentences, a context vocabulary with the corresponding term-context matrix. Each vector element represents the raw co-occurrence frequency between target words and context features using sentence boundaries as the context window.

2.8 Summary

This chapter presented methods for modelling document collections using statistical approaches. Topic models are statistical methods that summarise the content of document collections into a set of latent variables called topics. In addition, types of popular topic models have been described, i.e. LSA, pLSA, LDA and CTM.

Topic models have appealing characteristics for organising and visualising document collections. Topics can be used to cluster documents that share a common subject. A variety of interfaces have been developed for that purpose, however these systems usually do not deal with well-known problems of topic models. We presented previous work on tackling these three problems: (1) identifying incoherent topics, (2) identifying similar topics, and (3) representing topics using an alternative representation than lists of keywords.

Finally, we described distributional semantic representations of words which will be later used in the thesis for computing topic coherence and similarity.

EVALUATING TOPIC COHERENCE

As presented in Section 2.4, topic models can output topics that are difficult to interpret. Table 3.1 shows a sample of incoherent topics identified by humans in three different domains. Providing humans with these topics in a search interface can lead to a poor browsing experience. Filtering out these topics can improve browsing interfaces for exploring the content of document collections. In addition, identifying incoherent topics can be a useful pre-processing step for topic labelling algorithms (Lau et al., 2011) (see Section 2.6). Intuitively, labelling algorithms would fail at assigning labels to topics that do not represent a coherent thematic subject. Measuring topic coherence can also be utilised in topic model evaluation. Chang et al. showed that humans find topics generated by models with high predictive likelihood to be less coherent than topics generated from others with lower predictive likelihood. Following Chang’s findings, recent work on evaluation of topic models has been focused on automatically measuring the coherence of generated topics by comparing them against human judgements (Mimno et al., 2011; Newman et al., 2010b).

This chapter explores methods for automatically determining the coherence of topics. It proposes a novel approach for measuring topic coherence based on the distributional hypothesis which states that words with similar meanings tend to occur in similar context (Harris, 1954). Wikipedia is used as a reference corpus to create a distributional semantic model (Erk, 2012; Turney and Pantel, 2010). Each topic word is represented as a bag of highly co-occurring context

Domain	Topic
News	privacy, andrews, elli, alexander, burke, zoo, information, chung, user, regan
News	apple, evans, peru, portugal, ant, dinosaur, sherman, rent, portuguese, fossil
Email	umd, mathew, 800, adobe, mantis, quadra, wam, maryland, co, vram
Email	duke, event, expose, tyre, window, draw, den, p2, drawing, p1
Scientific	receptor, ohe, ry, cyp17, ryr2, ga, insp, korea, modification, binding
Scientific	eacute, france, germany, auml, uuml, dr, ouml, la, paris, hospital

Figure 3.1: A sample of less coherent topics generated by a topic model in three different domains (news articles, newsgroup emails and scientific articles). Topics are represented using the top- n most probable words.

words that are weighted using either the Pointwise Mutual Information (PMI) or a normalised version of PMI (NPMI). It also explores creating the vector space using differing numbers of context terms. All methods are evaluated by measuring correlation with human judgements on three different sets of topics. Results indicate that measures on the fuller vector space are comparable to the state-of-the-art proposed by Newman et al. (2010b), while performance consistently improves using a reduced vector space.

The chapter is organised into five sections. Section 3.1 describes the methodology of computing topic coherence using distributional semantics. Section 3.2 presents the experimental set-up while Section 3.3 discusses the results obtained. Section 3.4 presents a discussion of the results. Finally, Section 3.5 summarises the chapter.

3.1 Methodology

Previous work on identifying incoherent topics is based on computing the average pairwise similarity between topic words using word association measures (Mimno

et al., 2011; Newman et al., 2010b) (see Section 2.4 for a detailed description of previous work).

This thesis proposes a method for determining topic coherence based on computing distributional similarity between the top- n words in the topic. Following previous work, each topic is represented by a list of 10 words.

3.1.1 Topic Coherence

Let $T = \{w_1, w_2, \dots, w_n\}$ be the top- n most probable words from a topic generated from a topic model. Newman et al. (2010b) assume that the higher the average pairwise similarity between words in T , the more coherent the topic. Given a symmetric word similarity measure, $Sim(w_i, w_j)$, topic coherence is defined as follows:

$$Coherence_{sim}(T) = \frac{\sum_{\substack{1 \leq i \leq n-1 \\ i+1 \leq j \leq n}} Sim(w_i, w_j)}{\binom{n}{2}} \quad (3.1)$$

where $w_i, w_j \in T$.

3.1.2 Computing Topic Word Similarity

Each topic word is represented as a vector in a semantic space. Let $\vec{w}_1, \vec{w}_2, \dots, \vec{w}_n$ be the vectors which represent the top n most probable words in the topic. Also, assume that each vector consists of N elements and w_{ij} is the j th element of vector \vec{w}_i .

The similarity between the topic word vectors, and therefore coherence of the topic, is computed using the following standard measures used in work on distributional semantics (Curran, 2003; Grefenstette, 1994):

- The **cosine** of the angles between the vectors:

$$Sim_{cos}(\vec{w}_i, \vec{w}_j) = \frac{\vec{w}_i \cdot \vec{w}_j}{\|\vec{w}_i\| \|\vec{w}_j\|} \quad (3.2)$$

- The **Dice** coefficient:

$$Sim_{Dice}(w_i, w_j) = \frac{2 \times \sum_{k=1}^N \min(w_{ik}, w_{jk})}{\sum_{k=1}^N (w_{ik} + w_{jk})} \quad (3.3)$$

- The **Jaccard** coefficient:

$$Sim_{Jaccard}(w_i, w_j) = \frac{\sum_{k=1}^N \min(w_{ik}, w_{jk})}{\sum_{k=1}^N \max(w_{ik}, w_{jk})} \quad (3.4)$$

Each of these measures estimates the similarity between a pair of topic word vectors and can be substituted into equation 3.1 to produce a topic coherence measure based on distributional semantics.

Alternatively, the coherence of a set of topic words can be estimated with a single measure by computing the average similarity between the top- n topic word and the centroid:

$$Sim_{centroid} = \frac{\sum_{t \in T} sim_{cos}(T_c, t)}{n} \quad (3.5)$$

where T_c is the centroid of the vectors for topic T . For the experiments reported below the distance of each vector to the centroid is computed using the cosine measure¹.

3.1.3 Constructing the Semantic Space

Vectors representing the topic words are constructed from a semantic space consisting of information about word co-occurrence. The semantic space was created using Wikipedia² as a reference corpus and a window of ± 5 words³.

Weighting Vectors

Using the co-occurrence information to generate vectors directly does not produce good results. Therefore, the vectors are weighted using two approaches.

¹Experiments with Dice and Jaccard metrics produced lower performance.

²<http://dumps.wikimedia.org/enwiki/20120104/>

³Experiments with different lengths of context windows produced lower performance.

In the first, **PMI**, the pointwise mutual information for each term in the context is used rather than the raw co-occurrence count. Note that this application of PMI for topic coherence is different from one previously reported by Newman et al. (2010b) since PMI is used to weight vectors rather than to compute a similarity score between pairs of words. In addition, vectors are also weighted using **NPMI** (Normalised PMI).

Finally, the parameter γ is used to assign more emphasis on context features with high PMI (or NPMI) values with a topic word. Vectors are weighted using $\text{PMI}(w_i, f_j)^\gamma$ or $\text{NPMI}(w_i, f_j)^\gamma$ where w_i is a topic word and f_j is a context feature. For all of the experiments reported here, γ is set at 2 which was found to produce the best results.

Reducing the Basis

Including all co-occurring terms in the vectors leads to a high dimensional space. A semantic space with a smaller basis is formed by experimenting with three approaches to reducing the number of terms. Firstly, a **Top-N Semantic Space** is created by choosing the N most frequent context features in the reference corpus. N is set to 5,000 which found to produce better estimates of topic coherence ⁴.

In addition, following Islam and Inkpen (2006), a **Reduced Semantic Space** is created by choosing the β_{w_i} most related context features for each topic word w_i :

$$\beta_{w_i} = (\log(c(w_i)))^2 \frac{\log_2(m)}{\delta} \quad (3.6)$$

where δ is a parameter for adjusting the number of features for each word and m is the size of the corpus. Varying the value of δ did not affect performance for values above 1. This parameter was set of 3 for the results reported here. In addition a frequency cut-off of 20 was also applied since it is known that PMI is biased towards low frequencies.

Finally, a smaller semantic space was created by considering only topic words as context features, leading to n features for each topic word. This is referred to as the **Topic Word Space**.

⁴Values of N between 1,000 and 10,000 were also tested producing consistently lower results.

3.2 Evaluation

This section describes the creation of a data set for evaluating topic coherence. It presents the textual data used to train topic models in different domains and how the gold standard annotations of topic coherence were obtained.

3.2.1 Data

Since there are no standard data sets available for evaluating topic coherence, one has been developed for this study and has been made publicly available⁵. Newman et al. (2010b) used a similar approach to constructing a data set to evaluate coherence but it is not publicly available. A total of 300 topics are generated by applying LDA (see Section 2.2.3) over three different document collections:

- **NYT:** 47,229 New York Times news articles published between May and December 2010 from the GigaWord corpus. A set of 200 topics were generated and 100 randomly selected.
- **20NG:** The 20 News Group Data Collection⁶ (20NG), a set of 20,000 newsgroup emails organised into 20 different subjects (e.g. sports, computers, politics). Each topic is associated with 1,000 documents. 100 topics were generated from this document collection.
- **Genomics:** 30,000 scientific articles published in 49 journals from MEDLINE, originally used in the TREC-Genomics Track⁷. 200 topics were generated and 100 randomly selected.

All documents were pre-processed by removing stop words and lemmatising. Topics are generated using *gensim*⁸ with hyperparameters (α, β) set to $\frac{1}{num_of_topics}$. Each topic is represented by its 10 most probable words, a common approach in the literature (Mimno et al., 2011; Newman et al., 2010b).

⁵The data set can be downloaded from <http://staffwww.dcs.shef.ac.uk/people/N.Aletras/resources/TopicCoherence300.tar.gz>

⁶<http://people.csail.mit.edu/jrennie/20Newsgroups>

⁷<http://ir.ohsu.edu/genomics>

⁸<http://radimrehurek.com/gensim>

Evaluation of Topic Quality

Instructions -

You will be presented with 10 word sets, each of which represents a topic. You should judge the quality of each topic on a scale from 3 (Useful) to 1 (Useless).

A "Useful" topic is one that is semantically coherent, meaningful and interpretable. The subject will also be clear and can easily be labeled. For example given the set [Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday], we can say that it is "Useful" because one can easily infer the subjects "Week" or "Days of the Week".

If some of the topic words are coherent and interpretable but others are not then the topic would be judged as "Average" (denoted as 2 in the quality scale).

Finally, if the words appear random and unrelated to each other then the topic is judged as "Useless" (denoted as 1 in the quality scale). For example a set of random numbers, random characters or unrelated words.

Your answers will be entirely confidential and anonymous, and will be used for research purposes only. In accordance with ethical guidance, all information you provide will remain entirely anonymous. You are free to withdraw at any time by simply closing the web browser.

The survey is conducted in the context of the EU "Paths Project". The survey is undertaken by Nikolaos Aletras under the supervision of Dr. Mark Stevenson in the University of Sheffield.

Contact: n.aletras@dcs.shef.ac.uk

storm, weather, wind, temperature, rain, snow, air, high, cold, northern

Rating

Useless	1	2	3	Useful
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

police, officer, crime, street, man, city, gang, suspect, arrested, violence

Rating

Useless	1	2	3	Useful
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

oil, louisiana, coast, gulf, orleans, spill, state, fisherman, fishing, seafood

Rating

Useless	1	2	3	Useful
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

Figure 3.2: A screenshot of the user survey for collecting human judgements of topic coherence.

3.2.2 Human Evaluation of Topic Coherence

Human judgements of topic coherence were collected through a crowdsourcing platform, CrowdFlower⁹. Participants were presented with lists of the top-10 topic keywords. They were asked to judge topic coherence on a 3-point Likert scale from 1-3, where 1 denotes a "Useless" topic (i.e. words appear random and unrelated to each other), 2 denotes "Average" quality (i.e. some of the topic words

⁹<http://crowdflower.com>

are coherent and interpretable but others are not), and 3 denotes a “Useful” topic (i.e. one that is semantically coherent, meaningful and interpretable). Figure 3.2 shows a screen shot of the survey. Each participant was asked to judge up to 100 topics from a single collection. The average response for each topic was calculated as the coherence score for the gold-standard.

To ensure reliability and avoid random answers, the survey included a number of questions with predefined answers (Kazai, 2011) (either totally random words as topics or obvious topics such as week days). Annotations from participants that failed to answer these questions correctly were removed.

Three surveys were run, one for each collection of 100 topics. 1,778 filtered responses from 26 participants were obtained for the NYT data set and 1,707 from 24 participants for the 20NG data set. Participants were recruited by a broadcast email sent to all academic staff and graduate students in the University of Sheffield. For the Genomics data set the emails were sent only to members of the medical school and biomedical engineering departments. A total of 1,050 judgements from 12 participants were collected for this data set.

Inter-annotator agreement (IAA) is measured as the average of the Spearman correlation between the set of scores of each survey respondent and the average of the other respondents’ scores. The IAA in the three surveys is 0.70, 0.64 and 0.54 for NYT, 20NG and Genomics respectively.

3.2.3 Evaluation Metric

Performance is measured as the correlation between the similarity scores returned by each proposed method and the human judgements. This approach has been used to evaluate similar tasks including word and text similarity, e.g. (Agirre et al., 2009, 2012; Budanitsky and Hirst, 2001). In our experiments, we make use of the Spearman’s correlation coefficient. We prefer to use Spearman’s rather than Pearson’s correlation coefficient since it does not assume the relationship to be linear.

	NYT	20NG	Genomics			
Newman et al. (2010b)	0.71	0.73	0.73			
Average NPMI	0.74	0.76	0.76			
Mimno et al. (2011)	-0.39	0.34	-0.40			
Top-N Semantic Space						
	PMI	NPMI	PMI	NPMI	PMI	NPMI
Cosine	0.52	0.51	0.66	0.66	0.54	0.53
Dice	0.50	0.50	0.65	0.65	0.51	0.52
Jaccard	0.51	0.50	0.65	0.66	0.50	0.52
Centroid	0.52	0.52	0.66	0.66	0.53	0.53
Reduced Semantic Space						
	PMI	NPMI	PMI	NPMI	PMI	NPMI
Cosine	0.69	0.68	0.78	0.79	0.74	0.73
Dice	0.63	0.62	0.77	0.78	0.69	0.68
Jaccard	0.63	0.61	0.77	0.78	0.69	0.76
Centroid	0.67	0.67	0.77	0.78	0.73	0.71
Topic Words Space						
	PMI	NPMI	PMI	NPMI	PMI	NPMI
Cosine	0.76	0.75	0.79	0.80	0.80	0.80
Dice	0.68	0.71	0.79	0.80	0.79	0.80
Jaccard	0.69	0.72	0.80	0.80	0.80	0.80
Centroid	0.76	0.75	0.78	0.79	0.80	0.80

Table 3.1: Performance of methods for measuring topic coherence (Spearman Rank correlation with human judgements).

3.3 Results

Table 3.1 shows the results obtained for all of the methods on the three data sets. Performance of each method is measured as the average Spearman correlation with human judgements. The top row of each table shows the result using the average PMI approach (Newman et al., 2010b) while the next two rows show the results obtained by substituting PMI with NPMI and the method proposed by Mimno et al. (2011) (see Section 2.4). The main part of each table shows performance using the approaches described in Section 3.1 using various combinations of methods for constructing the semantic space and determining the similarity between vectors.

Using the average PMI between topic words correlates well with human judgements, 0.71 for NYT, 0.73 for 20NG and 0.75 for Genomics confirming results reported by Newman et al. (2010b). NPMI performs better than PMI, with an improvement in correlation of 0.03 for all data sets. The improvement is down to the fact that NPMI reduces the impact of low frequency counts in word co-occurrences and therefore creates more reliable estimates (Bouma, 2009).

On the other hand, the method proposed by Mimno et al. (2011) does not correlate well with human judgements and has the lowest performance of all of the methods tested (-0.39 for NYT, 0.34 for 20NG and -0.4 for Genomics). This demonstrates that while co-document frequency helps to generate more coherent topics (Mimno et al., 2011), it is not as reliable as word co-occurrence in a larger reference corpus. Lau et al. (2014) also confirms that this method does not work well.

Results obtained using the Top-N semantic space and the reduced semantic space and PMI are lower than the average PMI and NPMI approaches for the NYT and Genomics data sets. For the 20NG data set the results are higher than the average PMI and NPMI using these approaches. The difference in relative performance is down to the nature of these corpora. The words found in topics in the NYT and Genomics data sets are often polysemous or collocate with terms which become context features. For example, one of the top context features of the word “coast” is “ivory” (from the country). However, that feature does not exist for terms that are related to “coast”, such as “beach” or “sea”. The majority of

Topic Terms	Human Rating
Top-3	
family, wife, died, son, father, daughter, life, became, mother, born	2.63
election, vote, voter, ballot, state, candidate, voting, percent, party, result	3
show, television, tv, news, network, medium, fox, cable, channel, series	2.82
Bottom-3	
lennon, circus, rum, whiskey, lombardi, spirits, ranch, idol, make, vineyard	1.93
privacy, andrews, elli, alexander, burke, zoo, information, chung, user, regan	1.25
twitter, board, tweet, followers, conroy, halloween, kay, hands, emi, post	1.53

Table 3.2: Top-3 and bottom-3 ranked topics using Topic Word Space in NYT together with human ratings.

topics generated from 20NG contain meaningless terms due to the noisy nature of the data set (emails) but these do not suffer from the same problems with ambiguity and prove to be useful for comparing meaning when formed into the semantic space.

Similar results are obtained for the reduced semantic space using NPMI as the association measure. Results in NYT and Genomics are often around 0.01 lower than PMI while they are 0.01 higher for all of the methods in 20NG. This demonstrates that weighting co-occurrence vectors using NPMI produces little improvement over using PMI, despite the fact NPMI has better performance when the average similarity between each pair of topic terms is computed.

When the topic word space is used there is a consistent improvement in performance compared to the average PMI (Newman et al., 2010b) and NPMI approaches. More specifically, cosine similarity using PMI is consistently higher (0.05-0.06) than average PMI for all data sets and 0.02 to 0.04 higher than average NPMI (0.76, 0.79, 0.8 for NYT, 20NG and Genomics respectively). One

reason for this improvement in performance is that the noise caused by polysemy and high dimensionality of the context features of the topic words is reduced. Moreover, cosine similarity scores in the reduced semantic space are higher than average PMI and NPMI in all of the data sets, demonstrating that vector-based representation of the topic words produces better results than computing their average relatedness.

Another interesting finding is that the cosine metric produces better estimates of topic coherence compared to Dice and Jaccard in the majority of cases, with the exception of 20NG in reduced semantic space using PMI. Furthermore, similarity to the topic centroid achieves performance comparable to cosine.

3.4 Discussion

Table 3.2 shows the top-3 and bottom-3 ranked topics using Topic Word Space in NYT together with human ratings. The Top-3 topics represent concrete themes, i.e. family relations, election, TV. On the other hand, the interpretation of the Bottom-3 topics is non-trivial. These topics consist of mixtures of words usually without any semantic relatedness between each other, i.e. burke and zoo in the second topic.

The methods proposed in this chapter, i.e. Topic Word Space and average NPMI, produce reliable estimates of topic coherence since they tackle well-known problems of distributional semantics which is the high dimensionality of these spaces and the bias of PMI towards low frequencies. Lau et al. (2014) confirmed the effectiveness of our proposed methods showing that they perform best in two tasks: word intrusion (Chang et al.) and observed coherence.

3.5 Summary

This chapter proposed the use of distributional semantic similarity methods for automatically measuring the coherence of sets of words generated by topic models. Representing topic words as vectors of context features and then applying similarity metrics on vectors was found to produce reliable estimates of topic co-

herence. In particular, using a semantic space that consisted of only the topic words as context features produced the best results and consistently outperforms previously proposed methods for the task.

The method based on the topic word space which produced the best results is modified to measure topic similarity (in Chapter 4). It is also employed in Chapters 5, 6 and 7 to filter-out incoherent topics as a pre-processing step in topic labelling algorithms.

MEASURING TOPIC SIMILARITY

Some topics generated by a topic model will be similar while others are dis-similar. For example, a topic about basketball (TEAM GAME JAMES SEASON PLAYER NBA PLAY KNICKS COACH LEAGUE)¹ is more similar to topics about football (WORLD CUP TEAM SOCCER AFRICA PLAYER SOUTH GAME MATCH GOAL), or golf (GOLF, WOODS, HOLE, OPEN, COURSE, SHOT, ROUND, TOUR, PLAYER, TH) than one about the global finance (FED FINANCIAL BANKS FEDERAL RESERVE BANK BERNANKE RULE CRISIS CREDIT).

Methods that can automatically determine the similarity between topics would assist in the comprehension of topic models. For example, they could be applied within topic browsers by identifying related topics that could be clustered together or to provide links to similar topics (see Section 2.3).

LDA (see Section 2.2.3) cannot capture such correlations unless the semantic similarity between topics is measured. On the other hand, other topic models, e.g. CTM (see Section 2.2.4), have been introduced to identify correlations between topics and overcome this limitation of LDA. In CTM, the distribution over topics for a document is drawn from a logistic normal distribution. The covariance matrix for parametrising the logistic normal distribution can be used to identify correlations between topics and form a topic graph in which each node represents a topic and each edge denotes the correlation between them.

This chapter explores methods for measuring semantic similarity between top-

¹Topics are represented here using the 10 keywords with the highest marginal probabilities.

ics. This can be thought of as a post-processing step in LDA. In CTM, it can be viewed as re-writing the topic graph. Passos et al. (2011) showed that automatically generated topics often contain polysemous words which are assigned with high probabilities across many topics resulting in spurious correlations.

Approaches to computing topic similarity have been described in the literature but they have been restricted to using information from the word probability distribution to compare topics and have not been directly evaluated (see Section 2.5). The work in this chapter addresses these limitations by providing a systematic evaluation of a range of approaches to computing similarity between topics. Its contributions are to: (1) propose approaches for measuring topic similarity that rely on distributional semantics; (2) introduce a data set consisting of pairs of topics together with human judgements of similarity; (3) evaluate the proposed approaches using this data set; and (4) demonstrate that methods proposed here perform better than those used previously.

This chapter consists of four sections. First, Section 4.1 describes various methods of computing topic similarity. The second, Section 4.2 presents the experimental set-up and evaluation while Section 4.3 discusses the results obtained. Finally, Section 4.4 presents a summary of the chapter.

4.1 Methodology

A broad range of approaches for measuring similarity between topics are compared. We begin by applying measures based on the topics' word probability distributions which have been described in the literature (see Section 4.1.1). We also explore methods that make use of distributional similarity measures applied over semantic spaces produced from the topic model itself (Section 4.1.2), by measuring co-occurrences of words in a reference corpus (Section 4.1.3) and from the training corpus (Section 4.1.4). Three knowledge-based methods (Section 4.1.5) and a combination of approaches (Section 4.1.6) are also applied.

4.1.1 Topic Word Probability Distribution Similarity

We first experimented with topic similarity measures based on comparison of the topics' word distributions. We applied the JSD, KL-divergence and Cosine approaches² (see Section 2.5) and the Log Odds Ratio used by Chaney and Blei (2012).

4.1.2 Topic Model Semantic Space

The semantic space generated by the topic model can be used to represent the topics and the topic words (see Section 2.2). By definition each topic is a probability distribution over the words in the training corpus. In addition, each topic word can be represented as a vector with topics as features weighted by the probability of the word in each topic. For a corpus with D documents and W words, a topic model learns a relation between words and topics, T , and a relation between topics and documents as:

- a $T \times W$ matrix, Φ , that indicates the probability of each word in each topic, and
- a $D \times T$ matrix, Θ , that indicates the probability of each topic in each document.

We assume that Φ is the topic model semantic space and each topic word can be represented as a vector, Φ_i in Φ , with topics as features weighted by the probability of the word in each topic. Then, the similarity between two topics is computed as the average pairwise cosine similarity between their top-10 most probable words (**TS-Cos**).

4.1.3 Reference Corpus Semantic Space

Alternatively, topic words can be represented as vectors in a semantic space constructed using an external source. We adapt the method proposed in Section 3.1 for measuring topic similarity using distributional semantics. We make use of

²We also experimented with the other approaches described by Kim and Oh (2011) but found they did not perform well and do not report the results here.

Wikipedia as a reference corpus to count word co-occurrences and frequencies using a context window of ± 10 words centred on a topic word. The main advantage of using Wikipedia is that it is general and large enough to cover a broad range of thematic subjects.

Top-N Features

A semantic space is constructed considering only the top n most frequent words in Wikipedia (excluding stop words) as context features. Each topic word is represented as a vector of n features weighted by computing the Pointwise Mutual Information (PMI) between the topic word and each context feature, $\text{PMI}(w_i, w_j)^\gamma$. γ is a variable for assigning more importance to higher PMI values. In our experiments, we set $\gamma = 3$ and found that the best performance is obtained for $n = 5000$. Similarity between two topics is defined as the average cosine similarity of the topic word vectors (**RCS-Cos-N**).

Topic Word Space

Alternatively, we consider only the top-10 topic words from the two topics as context features to generate topic word vectors. Then, topic similarity is computed as the pairwise cosine similarity of the topic word vectors (**RCS-Cos-TWS**) similar to the approach described above (Section 4.1.2).

Word Association

Topic similarity can also be computed by applying word association measures directly. Newman et al. (2010b) measure topic coherence as the average PMI between the topic words. This approach can be adapted to measure topic similarity by computing the average pairwise PMI between the topic words in a pair of topics (**PMI**).

4.1.4 Training Corpus Semantic Space

We also experiment with using the training corpus to create semantic spaces. We create a term-document matrix such that each term in the vocabulary is

represented as a vector of documents in the corpus (Turney and Pantel, 2010). The values in these vectors are positive if the term is found in the document and 0 otherwise. In addition, we experiment with creating a semantic space by collecting co-document frequencies in the training corpus following the method proposed by Mimno et al. (2011) for measuring topic coherence.

Term-Document Space

Let \mathbf{C} be a term-document matrix and suppose that our training corpus consists of N documents and M unique terms (see Section 2.1). The matrix \mathbf{C} has M rows and N columns. Each term (row) represents a topic word vector. Element c_{ij} in \mathbf{C} is the tf.idf of the term i in document j . Topic similarity is computed as the pairwise cosine similarity of the topic word vectors (**TCS-Cos-TD**).

Word Co-occurrence in Training Documents

Alternatively, we generate a matrix \mathbf{Z} of co-document frequencies. The matrix \mathbf{Z} consists of N rows and N columns representing the N vocabulary words. Element z_{ij} is the log of the number of documents that contains the words i and j normalised by the document frequency, DF, of the word j , i.e.

$$z_{ij} = \log \frac{\text{DF}(i, j) + 1}{\text{DF}(j)} \quad (4.1)$$

Mimno et al. (2011) introduced that metric to measure semantic similarity between two topic words, and therefore topic coherence. We adapted it to estimate topic similarity as follows (**Doc-Co-occ**):

$$\text{Sim}_{\text{co-occ}}(T_i, T_j) = \frac{1}{2} \left(\sum_{\substack{m \in T_i \\ n \in T_j}} z_{nm} + \sum_{\substack{n \in T_j \\ m \in T_i}} z_{mn} \right) \quad (4.2)$$

where z_{nm} is the log of the number of documents containing the words n and m in topic T_i and T_j respectively. This metric aggregates the co-document frequency of the words between two topics and it is symmetric.

4.1.5 Knowledge-based Methods

The various approaches based on distributional similarity described above were compared against three existing knowledge-based methods.

UKB

Agirre et al. (2009) apply the Personalized PageRank algorithm (Haveliwala et al., 2003) to a graph created from WordNet to compute lexical similarity (**UKB**). We use UKB to generate a probability distribution over WordNet synsets for each word in the vocabulary W of the topic model. Similarity between two topic words is computed by transforming these distributions into vectors and comparing them using the cosine metric. If a topic word does not appear in WordNet its similarity value to every other word is set to 0. Similarity between two topics is computed by measuring pairwise similarity between their top-10 topic words, for each, selecting the highest similarity score.

Wikipedia Link Vector Model (WLVM)

Milne and Witten (2008) introduced an algorithm that identifies Wikipedia articles which are likely to be relevant to a given text. We apply their method to associate each topic to a set of Wikipedia articles. Then, similarity between Wikipedia articles is measured using the Wikipedia Link Vector Model (**WLVM**) (Milne, 2007) which uses both the link structure and the article titles of Wikipedia. Each link is weighted by the probability of it occurring. Thus, the value of the weight w for a link $x \rightarrow y$ between articles x and y is:

$$w(x \rightarrow y) = |x \rightarrow y| \times \log \left(\sum_{z=1}^t \frac{t}{z \rightarrow y} \right) \quad (4.3)$$

where t is the total number of articles in Wikipedia. The similarity of articles is compared by forming vectors of the articles which are linked from them and

computing the cosine of their angle.

$$\vec{x} = (w(x \rightarrow l_1), w(x \rightarrow l_2), \dots, w(x \rightarrow l_n)) \quad (4.4)$$

$$\vec{y} = (w(y \rightarrow l_1), w(y \rightarrow l_2), \dots, w(y \rightarrow l_n)) \quad (4.5)$$

where x and y are two Wikipedia articles and $x \rightarrow l_i$ is a link from article x to article l_i . Since the topics have been mapped to Wikipedia articles, similarity between two topics is computed by measuring pairwise similarity between articles using WLVM, for each, selecting the highest similarity score.

Explicit Semantic Analysis (ESA)

Explicit Semantic Analysis (**ESA**) (Gabrilovich and Markovitch, 2007) transforms the keywords of the topic into vectors that consist of Wikipedia article titles weighted by their relevance to the keyword. For each topic, the centroid is computed from the keyword vectors. Similarity between topics is computed as the cosine of the angle between the ESA centroid vectors.

4.1.6 Feature Combination Using Support Vector Regression

We also evaluate the performance of a support vector regression system (**SVR**) (Vapnik, 1998) with a linear kernel using a combination of approaches described above as features. With the exception of JSD, features based on the topics' word probability distributions were not used by SVR since it was found that including them reduced performance. All other approaches were included as features. The system is trained and tested using 10-fold cross validation.

4.2 Evaluation

This section presents the experimental set up for evaluating the proposed approaches to computing similarity between topics. First, we begin by creating a data set appropriate for the study since, to our knowledge, no standard data sets are available. The data set consists of pairs of topics generated by two topic

models (LDA and CTM) over two document collections using different numbers of topics. Secondly, the proposed approaches are evaluated by measuring the correlation with human similarity judgements obtained through crowdsourcing.

4.2.1 Data

We create two document collections. The first consists of 47,229 news articles from New York Times (NYT) in the GigaWord corpus and the second contains 50,000 articles from ukWAC (Baroni et al., 2009). We expect that ukWAC is more diverse than NYT since it contains random web pages while NYT contains news articles which have concise style of writing and content. Each article is tokenised then stop words and words appearing fewer than five times in the corpora removed. This results in a total of 57,651 unique tokens for the NYT corpus and 72,672 for ukWAC.

LDA Topics are learned by training LDA models over the two corpora using *gensim*³. The number of topics is set to $T = 50, 100, 200$ and hyperparameters, α and β , are set to $\frac{1}{T}$ where T is the number of topics.

CTM In addition, we make use of the C implementation provided by David Blei⁴ to train CTM using the EM algorithm. The number of topics to learn is set to $T = 50, 100, 200$ and the rest of the settings are set to their default values.

Incoherent topics are removed from each set of topics using the approach described in Section 3.1 using the Topic Words semantic space. Each topic is represented using the top 10 words with the highest marginal probability within it.

4.2.2 Generating Pairs of Topics

LDA Intuitively, each topic in a collection is likely to be similar to a small set of other topics. Randomly selecting pairs of topics will result in a data set in which the majority of pairs would not be similar. We overcome that problem

³<http://radimrehurek.com/gensim>

⁴<http://www.cs.princeton.edu/~blei/ctm-c/index.html>

Pair	Rating
obama, president, white, house, administration, bush, washington, barack, clinton, adviser, republican, democrat, house, senate, vote, obama, bill, democratic, congress, party	3.6
school, student, college, university, high, education, class, program, state, campus school, teacher, district, education, charter, union, system, state, public, turner	3.4
company, business, deal, executive, billion, million, share, firm, chief, credit gm, billion, stock, company, government, offering, share, motor, chrysler, general	2.4
team, game, james, season, player, nba, play, knicks, coach, league team, league, player, manager, baseball, bee, national, strasburg, major, cricket	2.0
tobacco, smoking, obesity, cigarette, health, soda, tax, smoker, snake, chemical hospital, health, care, massachusetts, state, medical, boston, patient, nonprofit, doctor	1.33
india, indian, delhi, country, government, mumbai, state, singh, world, company china, chinese, trade, state, united, country, export, currency, world, beijing	1.0
prize, nobel, liu, chinese, china, peace, award, right, committee, beijing investigation, official, department, state, agent, police, authorities, case, arrest, report	0.0
million, estate, money, gold, greene, ticket, real, rich, tax, dollar sport, race, marathon, bike, athlete, run, world, runner, mile, team	0.0

Table 4.1: A sample of 8 pairs of topics together with average humans judgements. Topics are represented by top-10 most probable words.

by assuming that the JSD between likely relevant pairs will be low while it will be higher for less relevant pairs of topics. We selected 800 pairs of topics out of

which 600 pairs consist of topics that have similar word distribution (in the top 6 most relevant topics of a given topic ranked by JSD). The other, 200 pairs are selected randomly. Table 4.1 shows a sample of pairs of topics together with their human judgements.

CTM The topic graph generated by CTM can be used to create all the possible pairs between topics that are connected in the graph rather than using JSD as described above. This results in a total of 70, 468 and 695 pairs in NYT, and a total of 80, 246 and 258 pairs in ukWac for the 50, 100 and 200 topics respectively.

4.2.3 Human Judgements of Topic Similarity

Human judgements of topic similarity were obtained using an online crowdsourcing platform, Crowdfunder⁵. Annotators were provided with pairs of topics and were asked to judge how similar the topics are and provide a rating on a scale of 0 (completely unrelated) to 5 (identical). Figure 4.1 shows a screen shot of the survey.

A set of control questions with obvious answers were also included in the survey to ensure reliability (Kazai, 2011). Annotations by subjects that failed to answer these questions correctly or gave the same rating for all pairs were removed.

The average response for each pair was calculated in order to create the final similarity judgement for use as a gold-standard. Inter-Annotator agreement (IAA) is computed by comparing each annotator against the average score generated by the four other annotators. The final IAA score is the average Spearman's correlation of these comparisons across all five annotators. The average IAA across all pairs for all of the collections is in the range of 0.53-0.68. The data set consisting of pairs of topics together with gold-standard annotations is freely available (<http://staffwww.dcs.shef.ac.uk/people/N.Aletras/resources/topicSim.tar.gz>).

⁵<http://crowdfunder.com>

Estimate the Similarity between Topics

Instructions

The aim of this survey is to collect information about how people judge the relatedness of topics in news articles. You will be presented with pairs of topics, represented by a set of 10 words each, and asked to judge how similar you think they are on the following scale:

- 5 - Identical
- 4 - Strongly Related
- 3 - Related
- 2 - Somewhat Related
- 1 - Unrelated
- 0 - Completely Unrelated

Topic 1: percent job month rate report sales economy increase quarter price
Topic 2: store sales customer retailer holiday brand price consumer retail business

General Similarity

	0	1	2	3	4	5	
Completely Unrelated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Identical

Topic 1: dance ballet dancer swan company dancing nutcracker balanchine ballerina choreographer
Topic 2: sport race marathon bike athlete run world runner mile team

General Similarity

	0	1	2	3	4	5	
Completely Unrelated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Identical

Topic 1: black white american african civil racial south race hispanic minority
Topic 2: center church religious christian community muslim religion mosque american group

General Similarity

	0	1	2	3	4	5	
Completely Unrelated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Identical

Figure 4.1: A screenshot of the user survey for collecting human judgements of topic similarity.

4.2.4 Evaluation Metric

Performance is measured as the correlation between the similarity scores returned by each proposed method and the human judgements using Spearman’s correlation coefficient.

4.2.5 Baseline

A simple **Word Overlap** baseline which measures the number of terms that two topics have in common normalised by the total number of topic terms (10

keywords for each topic) was also implemented.

4.3 Results

Tables 4.2 and 4.3 show the correlation between the topic similarity metrics described in Section 4.1 and average human judgements for the LDA and CTM topic pairs.

We begin by discussing the results obtained using the **Topics' Word Probability Distributions** (see Section 4.1.1). The correlations obtained using JSD, KL-divergence and Cos between topics are comparable with the baseline for the LDA data set for all of the topic collections and topic models. The metric proposed by Chaney and Blei (2012) is also based on comparison of word probability distributions and fails to perform well on either data set. These results suggest that the probability distribution metrics may be sensitive to the high dimensionality of the vocabulary. These metrics can also assign high similarity to topics that contain ambiguous words, resulting in low correlations with human judgements. For example a topic about the golf champion, Tiger Woods, could be identified as similar to a topic about forests. These results show that metrics based on word probability distributions that have previously been used to identify similar topics (see Section 2.5) do not perform well on this task and other methods should be explored.

Performance of the cosine of the word vector (TS-Cos) in the **Topic Model Semantic Space** (see Section 4.1.2) varies across different number of topics implying that the quality of the latent space generated by LDA and CTM is sensitive to this parameter.

The similarity metrics that make use of the **Reference Corpus Semantic Space** (see Section 4.1.3) consistently produce good correlations for topic pairs generated using both LDA and CTM. The best overall correlation for a single feature in most of the cases is obtained by the average PMI (in a range of 0.43-0.74). The performance of the distributional semantic metric using the Topic Word Space (RCS-Cos-TWS) is comparable and slightly lower for the top-N features (RCS-Cos-N). This indicates that the reference corpus covers a broader range of semantic subjects than the latent space produced by the topic model

Spearman's r						
	LDA					
	NYT			ukWAC		
Method	50	100	200	50	100	200
Baseline						
Word Overlap	0.32	0.40	0.51	0.22	0.32	0.41
Topic Word Probability Distribution						
JSD	0.37	0.44	0.53	0.29	0.30	0.34
KL-Divergence	0.29	0.29	0.41	0.20	0.24	0.33
Cos	0.31	0.37	0.59	0.30	0.30	0.36
Chaney and Blei (2012)	0.16	0.26	0.18	0.29	0.21	0.25
Topic Model Semantic Space						
TS-Cos	0.35	0.41	0.67	0.29	0.35	0.42
Reference Corpus Semantic Space						
RCS-Cos-N	0.37	0.46	0.61	0.35	0.32	0.39
RCS-Cos-TWS	0.40	0.54	0.70	0.38	0.43	0.51
PMI	<u>0.43</u>	<u>0.63</u>	<u>0.74</u>	0.43	0.53	<u>0.64</u>
Training Corpus Semantic Space						
TCS-Cos-TD	0.36	0.42	0.67	0.29	0.31	0.40
Doc-Co-occ	0.28	0.29	0.45	0.28	0.22	0.30
Knowledge-based						
UKB	0.25	0.38	0.56	0.22	0.35	0.41
WLVM	0.35	0.51	0.51	0.35	0.46	0.53
ESA	<u>0.43</u>	0.58	0.71	0.46	<u>0.55</u>	0.61
Feature Combination						
SVR	0.46	0.64	0.75	0.46	0.58	0.66
IAA	0.54	0.58	0.61	0.53	0.56	0.60

Table 4.2: Results for various approaches to topic similarity in LDA. All correlations are significant $p < 0.001$. Underlined scores denote best performance of a single feature. Bold font denotes best overall performance.

Spearman's r						
Method	CTM					
	NYT			ukWAC		
	50	100	200	50	100	200
Baseline						
Word Overlap	0.56	0.45	0.49	0.35	0.33	0.53
Topic Word Probability Distribution						
JSD	0.59	0.43	0.49	0.38	0.34	0.60
KL-Divergence	0.54	0.39	0.56	0.31	0.29	0.47
Cos	0.58	0.45	0.52	0.50	0.40	0.58
Chaney and Blei (2012)	0.29	0.40	0.31	-0.23	0.12	0.61
Topic Model Semantic Space						
TS-Cos	0.67	0.51	0.49	0.51	0.42	0.42
Reference Corpus Semantic Space						
RCS-Cos-N	0.60	0.47	0.61	0.57	0.42	0.41
RCS-Cos-TWS	0.63	0.59	0.62	0.60	0.55	0.54
PMI	0.68	<u>0.70</u>	0.64	0.58	<u>0.62</u>	<u>0.64</u>
Training Corpus Semantic Space						
TCS-Cos-TD	0.64	0.54	0.58	0.49	0.43	0.43
Doc-Co-occ	0.65	0.36	0.57	0.31	0.26	0.34
Knowledge-based						
UKB	0.52	0.41	0.40	0.41	0.43	0.42
WLVM	0.65	0.56	0.54	0.45	0.53	0.53
ESA	<u>0.69</u>	0.67	0.64	0.70	<u>0.62</u>	0.61
Feature Combination						
SVR	0.72	0.71	0.62	0.60	0.65	0.66
IAA	0.68	0.68	0.64	0.67	0.63	0.64

Table 4.3: Results for various approaches to topic similarity in CTM. All correlations are significant $p < 0.001$. Underlined scores denote best performance of a single feature. Bold font denotes best overall performance.

and therefore provides better semantic representations for the topic words and reliable similarity estimations.

When the term-document matrix from the **Training Corpus Semantic Space** (see Section 4.1.4) is used performance is worse than when the reference corpus is used. In addition, using co-document frequency derived from the training corpus does not correlate particularly well with human judgements. These methods are sensitive to the size of the corpus, which may be too small to generate reliable estimates of tf.idf or co-document frequency.

Correlations for the **Knowledge-based methods** (see Section 4.1.5) are good for the Wikipedia-based methods, WLVM and ESA. The WordNet-based metric, UKB, does not perform particularly well. The reason for the poor performance of UKB is that the topics often contain named entities that do not exist in WordNet (see the first pair of topics in Table 4.1). That does not happen with the Wikipedia metrics which perform consistently better. In particular ESA achieves performance comparable (or even better in some cases) to PMI. WLVM performs better than UKB but not as well as ESA. The reason for the lower performance of WLVM may be the Wikification algorithm, which is designed for coherent documents rather than lists of topic keywords.

Feature Combination using SVR gives the best overall result for LDA (in the range 0.46-0.75) and CTM (0.60-0.72). However, the feature combination performs slightly lower than the best single feature in two cases when CTM is used (T=200, NYT and T=50, ukWAC). Analysing the coefficients obtained by the SVR in each fold for these cases, we found that JSD and the Word Overlap reduce SVR performance. We repeated the experiments by removing these features which resulted in an improvement in correlation (0.64 and 0.65 respectively). However, these features seem to be quite useful in the rest of the experiments since a drop in SVR was observed when they are removed.

Another interesting observation is that the correlations of the various similarity metrics with human judgements increase with the number of topics in LDA for both corpora. This result is consistent with the findings of Stevens et al. (2012) that topic model coherence increases with the number of topics (between 10-200). Fewer topics makes the task of identifying similar topics in LDA more difficult because it is likely that they will contain some terms that do not relate to the

topic's main subject. Correlations in CTM are more stable for different numbers of topics because of the nature of the model. That is, the pairs have been generated using the topic graph which by definition contains correlated topics. On the other hand, the data sets for LDA are constructed by randomly sampling, as well as selecting pairs with low JSD.

4.4 Summary

This chapter explored the task of determining the similarity of automatically generated topics and described a range of approaches to the problem. Previous approaches to measuring similarity between topics have been based on comparison of topic's word probability distribution and have not been evaluated.

A wide range of approaches for measuring topic similarity have been proposed including distributional semantic metrics, based on the topic model space, a reference corpus and the training corpus, as well as existing knowledge-based methods.

Evaluation has been carried out on a data set of pairs of topics generated by two topic models, LDA and CTM, together with human judgements of similarity. The best performing metrics are those based on the reference corpus. In addition, a knowledge-based method based on Wikipedia, ESA, performs comparably to the reference corpus.

The most interesting finding is the poor performance of the metrics based on word probability distributions previously used for this task. The results obtained demonstrate that word association measures, such as PMI, and state-of-the-art textual similarity metrics, such as ESA, are more appropriate.

AUTOMATIC LABELLING OF TOPICS USING TEXT

Graph-based methods in NLP have been proposed to represent unstructured texts as graphs. The nodes of the graphs consist of the words in text while edges are usually weighted by computing the similarity of the words that connect. These methods have been proved to work well in document summarisation (Mihalcea and Tarau, 2004) by identifying important terms in text. Here, we make use of graph-based methods to identify important terms which intuitively may be indicative of longer keyphrases that summarise the main thematic subject of a topic. This chapter introduces a graph-based approach for labelling topics which is unsupervised and less computationally intensive than previous methods introduced for the task (see Section 2.6).

The proposed 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). Evaluation on a standard data set shows that the graph-based method consistently outperforms the best performing previously reported method, which is supervised (Lau et al., 2011).

The contributions of the work presented here are to: (1) introduce a graph-based method for selecting appropriate labels for automatically generated topics; (2) evaluate the proposed method on a standard dataset; (3) demonstrate that the proposed unsupervised method is consistently better than the best performing


```
{'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 5.1: Sample of the metadata associated with a search result.

previously reported method which is supervised.

This chapter is organised as follows. Section 5.1 introduces an unsupervised graph-based approach to topic labelling. Section 5.2 presents the experimental set-up. The results of the evaluation on a standard data set are presented in Section 5.3 and a summary of the chapter in Section 5.4.

5.1 Methodology

The proposed approach uses the top-N keywords for a topic to form a query that is submitted to a search engine. We assume that the results returned from this search are appropriate for the topic. They are analysed to identify the terms that are central to the topic. The suitability of candidate labels are evaluated based on terms extracted from these search results.

5.1.1 Generating Candidate Labels

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.

5.1.2 Retrieving and Processing Text Information

Information obtained from web searches is used to identify the best labels from the set of candidates. The top 10 topic keywords are used to form a query which is submitted to the Bing¹ search engine. Textual information included in the Title field² of the search results metadata was extracted. Each title was tokenised using openNLP³ and stop words removed.

Figure 5.1 shows a sample of the metadata associated with a search result for the topic: VMWARE, SERVER, VIRTUAL, ORACLE, UPDATE, VIRTUALIZATION, APPLICATION, INFRASTRUCTURE, MANAGEMENT, MICROSOFT. For example, the textual information extracted from that search result is “Microsoft | Server & Cloud | Datacenter | Virtualization ...”.

5.1.3 Creating a Text Graph

Remaining words in the search result are used to create a graph $G = (V, E)$. Each node, $v \in V$, 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 each edge, $e \in E$, has a weight $e = 1$. In addition, we weight the edges of the graph by computing the relatedness, $sim(v_i, v_j)$, between two nodes, v_i and v_j , using NPMI (see Equation 2.16).

Word co-occurrences are computed using Wikipedia as a corpus. Pairs of

¹<http://www.bing.com/>

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

³<http://opennlp.apache.org/>

words are connected with edges only if $\text{NPMI}(w_i, w_j) > 0.2$ thereby avoiding connections between words co-occurring by being added to the graph and introducing noise.

5.1.4 Identifying Important Terms

Important terms are identified by applying the PageRank algorithm (Page et al., 1999). PageRank was originally developed for assigning importance to set of web pages interconnected with hyperlinks. 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). The PageRank score (Pr) over G for a word (v_i) can be computed by the following equation:

$$Pr(v_i) = d \cdot \sum_{v_j \in C(v_i)} \frac{sim(v_i, v_j)}{\sum_{v_k \in C(v_j)} sim(v_j, v_k)} Pr(v_j) + (1 - d)\mathbf{v} \quad (5.1)$$

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.

5.1.5 Ranking Labels

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

$$Score(L) = \sum_{i=1}^m Pr(w_i) \quad (5.2)$$

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.

5.2 Evaluation

In this section, we describe the data, the evaluation framework and the model parameters used in our experiments.

5.2.1 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).

5.2.2 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**) (Croft et al., 2009; Järvelin and Kekäläinen, 2002) 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:

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i)} \quad (5.3)$$

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

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (5.4)$$

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

5.2.3 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 since including more results did not improve performance but required additional processing.

5.3 Results and Discussion

Results obtained for the various evaluation metrics are shown in Table 5.1. Performance obtained is shown when PageRank is applied to the unweighted (**PR**) and NPMI-weighted graphs (**PR-NPMI**) (see Section 5.1.3). 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

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 5.1: 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).

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 (see Section 2.6 for more details). 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 (**PR**) are consistently better than the supervised and unsupervised methods reported by Lau et al. (2011) for the Top-1 Average scores. This improvement compared to the unsupervised method is significant (t-test, $p < 0.05$) in all domains, while the improvement over the supervised method is only significant in PUBMED. A slight improvement in performance is observed when the weighted graph is used (**PR-NPMI**). 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 Lau et al.’s 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.

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 term importance. Consequently, it is not necessary to measure semantic similarity between topic keywords and candidate labels as previous approaches have done (see Section 2.6). 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.

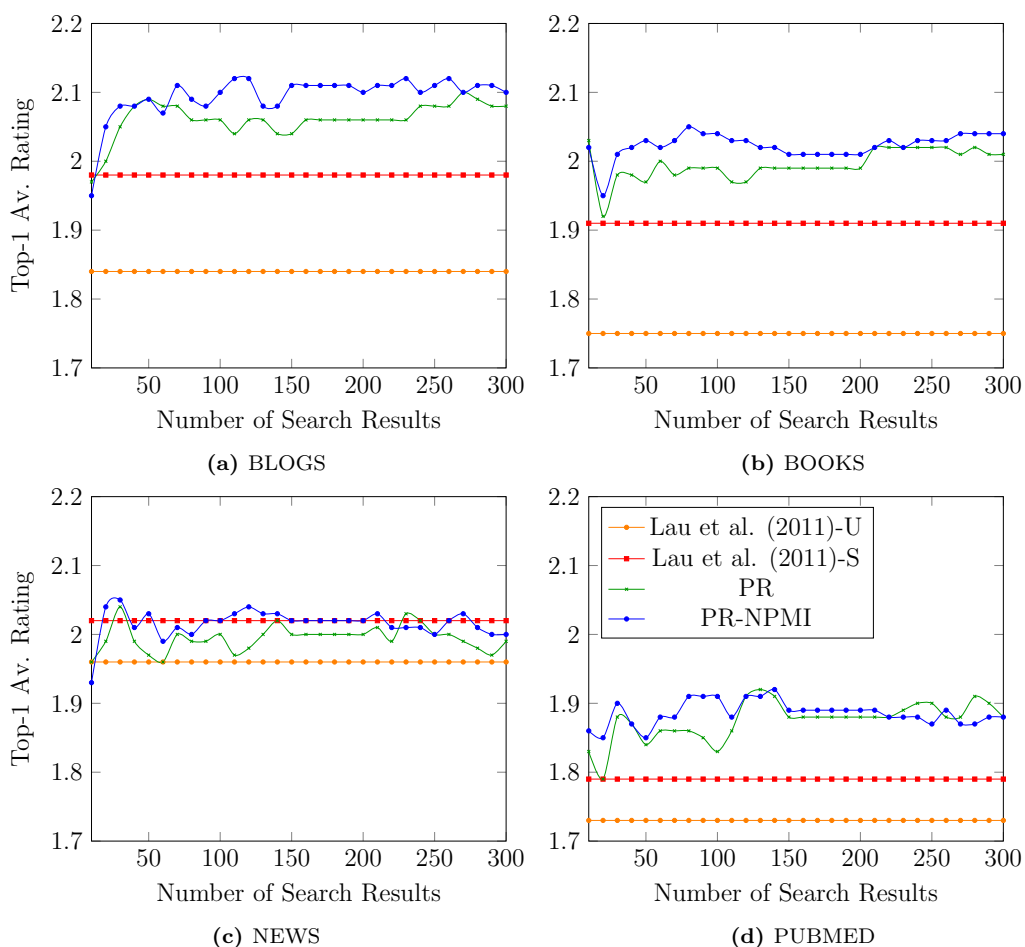


Figure 5.2: Top-1 Average Rating obtained for different number of search results.

5.3.1 Experimenting with the Number of Search Results

In Figure 5.2, 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 on good performance of ranking methods.

5.4 Summary

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.

AUTOMATICALLY LABELLING OF TOPICS USING IMAGES

The topic labelling techniques presented so far have focussed on the generation of textual labels (see Section 2.6 and Chapter 5). An alternative approach is to represent a topic using an illustrative image. Images have the advantage that they can be understood quickly. 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 (see Section 2.3). In addition, images are language independent and therefore can be used as an alternative to textual labels. This gives insights about the content of a text collection to people that are not familiar with the language of the text.

This chapter tackles the problem of identifying representative images that can be used to illustrate automatically generated topics. The proposed approach utilises the vast amount of pictures existing on the Web to generate a set of candidate images for each topic. Candidate images are retrieved 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. The proposed 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 outperforms three baselines.

This chapter consists of five sections. Section 6.1 describes the methodology for labelling topics using images. Section 6.2 describes the experimental set-up and evaluation while Section 6.3 presents the results obtained. Section 6.4 presents a discussion of the results. Finally, Section 6.5 summarises the chapter.

6.1 Methodology

This section describes an approach to identifying images to illustrate automatically generated topics. It is assumed that no candidate images are available so the first step (see Section 6.1.1) is to generate a set. However, in situations where a candidate set is available this first step is not required.

6.1.1 Selecting Candidate Images

The method presented here makes use of images from Wikipedia available under the Creative Commons licence, allowing it to be made publicly available. The top-5 terms¹ from a topic are used to query Google using its Custom Search API². The search is restricted to the English Wikipedia³ with image search enabled. The top-20 images retrieved for each search are used as candidates for the topic.

6.1.2 Feature Extraction

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

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 its content and (at least to some extent) related to the topic. The textual information is formed by

¹We noticed that in some cases the search engine does not return any results for longer queries.

²<https://developers.google.com/apis-explorer/#s/customsearch/v1>

³<http://en.wikipedia.org>

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 tokenising and removing stop words.

Visual Information

Image features represent important properties of an image. Major feature types are colour, shape, texture or salient points in an image and they are categorised as global and local features.

Global features tend to characterise an image as a whole. For example, the average of the intensities of red, green and blue colours gives an estimation of the overall colour distribution in the image. The main advantages of global features are the fast detection and extraction. However, they are not quite suitable to represent an image due to that they are sensitive to location (Datta et al., 2008). Examples of global features representations are colour histograms and global shape descriptors.

Local features capture interesting areas around certain pixels. For example an interesting area could be where colour intensity alternation between adjacent pixels is detected. Global description of an image is obtained by summarising local features (Datta et al., 2008). Local features are salient points and local shape descriptions in an image.

Nixon and Aguado (2008) define *low-level features* as basic features that can be extracted without taking into account any spatial information of an image. For example we may need to detect interesting locations in images such as object corners, boundary lines of an area or even, car wheels. These types of features are defined as keypoint features or interest points which are grouped together with neighbouring pixels and are described as patches or blobs (Szeliski, 2010). Another class of features are edges which lie within object boundaries.

The first step to represent images as sets of low-level features is to detect them. It has been shown that patches that contain large contrast changes (gradients) are easier to be detected by estimating local minima or maxima in their surfaces. Five popular detectors are: Harris points, Harris-Laplace regions, Hessian-Laplace regions, Harris-Affine regions and Hessian-Affine regions. For a detailed description

of patch and edge detectors refer to Nixon and Aguado (2008) and Szeliski (2010).

Features, i.e local patches, should be represented mathematically and invariantly to image transformations such as scale and rotation. Many different techniques have been developed such as distribution-based descriptors, spatial-frequency techniques and differential descriptors which are illustrated in Mikolajczyk and Schmid (2005); Szeliski (2010).

Scale Invariant Feature Transform (SIFT) (Lowe, 1999, 2004) is an approach for distinctive and invariant feature extraction and description. SIFT features are invariant to image scale and rotation and partially invariant to affine distortion, 3D viewpoint, noise and changes in illumination. The method includes stages from candidate location detection to its description.

Keypoint detection involves three stages. First, candidate locations are detected finding minima or maxima by searching the whole image over all scales using a difference-of-Gaussian function. Then, candidate locations are filtered by applying a threshold of minimum contrast. Finally, orientations are assigned to each keypoint.

Features are described by a 128-D vector. First, a keypoint is represented as a 16×16 grid of samples. For each sample, its weighted gradient magnitude and its orientation are computed producing 256 magnitude values. Then, these samples are added to gradient orientation histograms which are represented in a 4×4 quadrant containing 8 orientation bins. This results to 128 non-negative values for a feature description. Various descriptors are compared by Mikolajczyk and Schmid (2005) where SIFT performance in image matching exceeds performance of other reported methods.

Hence, we extract visual information from each image using low-level image keypoint descriptors, i.e. SIFT sensitive to colour information. Image features are extracted using dense sampling and described using Opponent colour SIFT descriptors provided by the *colordescrptor*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

⁴<http://koen.me/research/colordescrptors>

(BOVW).

6.1.3 Ranking Candidate Images

It is assumed that good illustrative images for a topic are ones that are similar to the others in the set of candidates and those with high similarity to the topic. Therefore, we experiment with graph-based algorithms for identifying image importance and measures of similarity between the topic and textual information associated with the candidate.

PageRank

PageRank is employed for identifying important images in a graph (see Section 5.1.3). 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 (similar to Equation 5.1):

$$Pr(V_i) = d \cdot \sum_{V_j \in C(V_i)} \frac{sim(V_i, V_j)}{\sum_{V_k \in C(V_j)} sim(V_j, V_k)} Pr(V_j) + (1 - d)\mathbf{v} \quad (6.1)$$

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.

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 6.1 to prefer certain nodes. The values in \mathbf{v} effectively initialises the graph and assigning high values to nodes in \mathbf{v} makes them more likely to be assigned a high

PPR score. PPR prefers images with textual information that is similar to the terms in the topic.

Weighting Graph Edges

Three approaches were compared for computing the values of $sim(V_i, V_j)$ 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 **PMI** (see Section 3.1.3). 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 **PR_{PMI}**. The second approach makes use of **ESA** (see Section 4.1.5) to create the graph and its PageRank is denoted as **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 **PR_{vis}** and it is similar to the approach proposed by Jing and Baluja (2008) for associating images to text queries.

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 $PR_{vis}+Per(PMI)$.

6.2 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 study⁵. Human judgements about the suitability of images are obtained through crowdsourcing.

6.2.1 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 categories (e.g. SPORTS, POLITICS, COMPUTING). Categories that have more than 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 using the *gensim* package⁶. The hyperparameters (α, β) were set to $\frac{1}{num_of_topics}$. Incoherent

⁵Data set can be downloaded from <http://staffwww.dcs.shef.ac.uk/people/N.Aletras/resources/datasetNAACL13.tar.gz>

⁶<http://pypi.python.org/pypi/gensim>

Is the Image Appropriate as a Label for the Topic?

Instructions ▾

You will be presented with sets of 10 words representing a topic together with an image that can be used to summarise the topic. You should score topic labels according to the following scale:


- 3 = very good label
- 2 = reasonable label
- 1 = somewhat related, but bad as a topic label
- 0 = completely inappropriate topic label

It is possible that there are no "very good" labels for some topics, so in some cases no labels will receive 3s. Random answers are identified, so please do not answer randomly.

You must score all labels in each page for the answers to be approved. Please make at least 160 annotations which is a number that makes your participation useful. Some topics are technical or domain-specific. If you are unsure about the meaning of a word, you may do a quick web look up.

Note: If you can see no image for a given topic you should wait to be loaded. If it doesn't load after 10 secs then you should rate with 0.

Topic: storm weather wind temperature rain snow air high cold northern
Image:



Rating

	0	1	2	3	
Inappropriate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Good

Figure 6.1: A screenshot of the user survey for collecting human judgements of image labels.

topics are filtered out by using the distributional semantics method (Topic Word Space) introduced in Chapter 3. 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 6.1.1, producing a total of 6,000 candidate images (20 for each topic).

6.2.2 Human Judgements of Image Relevance

Human judgements of the suitability of each image were obtained using Crowdflower (see Chapter 3 and 4). 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). Figure 6.1 shows a screenshot of the crowdsourcing experiment.

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.

The total number of filtered responses obtained was 62,221 from 273 participants. Each topic-image pair was rated by at least 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 6.2 shows three example topics from the data set together with the images that received the highest average score from the annotators. The scores assigned to the candidate images for some topics are higher than others. For example, the three candidate images for topics (1) and (2) in Figure 6.2 have scores in the range 2.83 to 2.73 while the highest score assigned to any of the candidate images for topic (3) is 1.91.

6.2.3 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. Two metrics are used: **Top-1 average rating** and **nDCG** (see Section 5.2.2).

6.2.4 Baselines

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

Topic (1): dance, ballet, dancer, swan, company, dancing, nutcracker, balanchine, ballerina, choreographer



2.8



2.8



2.73

Topic (2): wine, bottle, grape, flavor, dry, vineyard, curtis, winery, sweet, champagne



2.83



2.8



2.8

Topic (3): haiti, haitian, earthquake, paterson, jean, prince, governor, au, cholera, country



1.91



1.7



1.6

Figure 6.2: Sample of topics and the three image candidates that received the highest average annotation score (shown below image).

- **Average Human Ratings**

As we described above, each image label has been annotated by 10 humans and a gold standard score computed as the average of human judgements. The Average Human Ratings baseline is the average score from the 20 labels.

- **Word Overlap**

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. Similarity is defined as the number of terms shared by a topic and image candidate normalised by the sum of the terms in the topic and image’s metadata.

- **Google Image Search**

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

6.2.5 Human Performance

A study was conducted to estimate human performance on the image selection task. Three annotators (a staff member and two graduate students at our institution) 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. The average IAA across the three annotators was 0.59.

6.3 Results

Table 6.1 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 for the values 1, 3 and 5.

Model	Top-1 Av. Rat.	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
Human Performance	2.24	-	-	-

Table 6.1: Results for Various Approaches to Topic Labelling using Images.

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 outperform Random and Word Overlap 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 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 ($PR_{vis}+Per(PMI)$). The best results for the various DCG metrics are produced when both the graph and the personalisation vector are weighted using PMI scores ($PR_{PMI}+Per(PMI)$) while results for $PR_{vis}+Per(PMI)$ are comparable. 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. The main reason might be that PMI provides better similarity estimation between topic words and image metadata than ESA (see also Chapter 4). These results also indicate that graph-based methods for ranking images are useful for illustrating topics.

6.4 Discussion

Figure 6.3 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

dance, ballet, dancer, swan, company, dancing, nutcracker, balanchine, ballerina, choreographer



2.7



2.3



2.5

wine, bottle, grape, flavor, dry, vineyard, curtis, winery, sweet, champagne



2.7

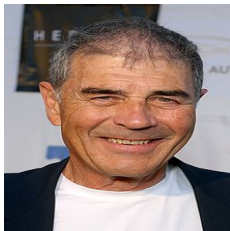


2.6



2.1

haiti, haitian, earthquake, paterson, jean, prince, governor, au, cholera, country



1.0



1.2



0.2

Figure 6.3: A sample of topics and their top-3 images selected by applying the $PR_{vis} + Per(PMI)$ approach. The number under each image represents its average human annotations score.

selected for the third topic do not seem to be as appropriate. Human judgements associated with the top-3 images selected for each topic confirm that. For the first two topics, the average human rating of the images is higher than 2 while for the third topic is below 1.

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” and pictures of families but no pictures of plants.

6.5 Summary

This chapter introduced the novel task of labelling topics using images and proposed an approach to selecting appropriate images. 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.

COMPARING TOPIC REPRESENTATIONS USING AN EXPLORATORY TASK

Previous chapters presented a range of topic labelling methods independently and without evaluation on a real application (see Chapter 5 and 6). 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 chapter aims to address, in carrying out a task-based evaluation of different topic model representations.

In this chapter, 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. 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 browser interface?
2. what is the impact of different topic representations on human search effec-

tiveness for a given query?

Section 7.1 introduces an experiment in which three approaches to topic labelling are applied and evaluated within an exploratory search interface. The results of the experiment are presented in Section 7.2 and conclusions in Section 7.3.

7.1 Methodology

We conducted a retrieval task to compare three topic representations: (1) lists of keywords (see Section 2.6), (2) textual labels (see Section 2.6 and Chapter 5), and (3) image labels (see Chapter 6).

7.1.1 Document Collection

We make use of a subset of the Reuters Corpus (Rose et al.) 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. Topic categories of Reuters Corpus are appropriate for the task since they cover a broad range of subjects (politics, sports, arts etc.). We selected 20 topic categories from which 100,000 documents extracted randomly.

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 7.1 shows the Reuters Corpus topic categories used to form the collection together with the number of associated documents.

7.1.2 Topic Modelling

We make use of the implementation provided by David Blei¹ to train an LDA model over the document collection using variational inference (Blei and Jordan,

¹<https://www.cs.princeton.edu/~blei/lda-c/index.html>

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

Table 7.1: Number of documents in each Reuters Corpus topic category

2003). 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 $T \geq 100$ (Stevens et al., 2012). Finally, we removed topics that are difficult to interpret to leave a total of 84 topics. Incoherent topics are filtered out by using the distributional semantics method (Topic Word Space) introduced in Chapter 3.


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

Table 7.2: Labels generated for an example topic.

7.1.3 Topic Browsing Systems

The topic browsing system developed for this study is based on the publicly available TMVE (Chaney and Blei, 2012) (see Section 2.3). We created three browsing systems. The three systems used different ways of representing topics: (1) keywords, (2) textual phrases and (3) images. By default the TMVE only supports keyword representation of topics, therefore we modified it to support textual and image labels. Table 7.2 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.

Keywords

Keywords are generated using the approach used by the TMVE, i.e. selecting the 10 keywords with the highest marginal probabilities for the topic (see Section 2.6).

Textual Labels

Textual labels are generated using a previously proposed approach by Lau et al. (2011) (see Section 2.6).

Time Remaining
02 32
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
- ai, egypt, jordan, king, egyptian, morocco, arab, libya, minister, hussein
- iraq, iraqi, military, gulf, baghdad, united, defence, saddam, kurdish, war
- france, french, paris, chircac, 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

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.


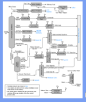


- modern, united states navy carrier air operations
- football league
- space shuttle
- budget
- french presidential election, 1995
- tourists
- security council
- morning
- floods
- united states department of state
- conference
- hun sen
- boris yeltsin
- collective security treaty organization
- south lebanon
- amnesty international
- companies
- fire department
- agreed framework
- romanian communist party

(b) Textual phrases

Time Remaining
02 30
MINUTES SECONDS

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 7.1: Topic browsing interfaces.

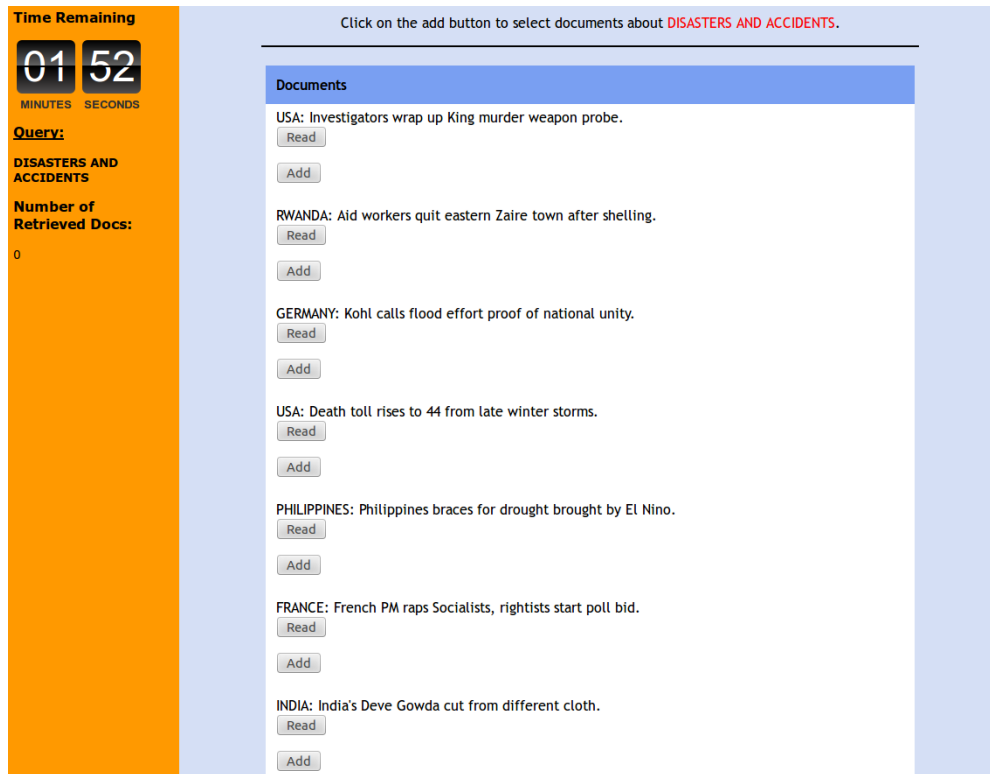


Figure 7.2: Topic browsing: List of documents.

Image Labels

We associate topics with image labels using the approach proposed in Chapter 6 by generating candidate labels from Wikipedia and ranking them using the $PR_{vis} + Per(PMI)$ approach.

7.1.4 Exploratory Search 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 7.1) 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. Participants are first provided with the list of LDA topics represented by one modality (keywords, text or image) and a query. They are asked to identify all topics that are potentially relevant to the query. Figure 7.1 shows the topic browser interface for the three different modalities. In the second step, the participant is presented with a list of documents associated with the topics selected. Documents are 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 7.2 shows a subset of the documents that are associated with the topics selected in the first step for the query *Disasters & Accidents*.

We also asked users to fill a post-task questionnaire once they had completed the retrieval task. The questionnaire consists of five questions which seek to give insights into participants' satisfaction with the retrieval task and the topic browsing systems. Participants had to assign a score from 1 to 7 in each question. First, we asked about the usefulness of topic representations, i.e. keywords, text 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 are as follows:

- How useful were the keywords to represent topics? (Usefulness (Keywords))
- How useful were the textual phrases to represent topics? (Usefulness (Text))
- 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)

7.1.5 Subjects and Procedure

We recruited 15 members of research staff and graduate students at Universities of Sheffield, Melbourne and King's College for the user study. All of the participants have a computer science background and were also all familiar with on-line digital library and retrieval systems.

Each participant first was 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 of the task, see how many queries they have completed so far or selecting to perform a new query.

Participants were asked to carry out each of the 20 queries in a random order. The topic representation for each query was randomly chosen and participants were asked to carry out queries using each of the three possible 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 are able to carry out 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 task within a week.

7.2 Results

7.2.1 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 7.3 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. On the other hand, the number of documents retrieved when topics are represented by textual labels is higher (1,264). This suggests 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 and lower than textual labels.

We also observed that the number of retrieved documents is high for queries

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

Table 7.3: Number of retrieved documents for each query and topic representation.

that are associated with many relevant documents (*Sports* in keywords, textual 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 distinct from other queries, making it easier to identify relevant documents. 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*.

We further examine 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. High correlation was observed between keywords and text (0.76) and keywords and image (0.74) while the correlation between text and image is lower (0.63). A possible reason for this might be that both textual and image labels are automatically generated which results in the introduction of noise. Comparing two noisy methods has a lower correlation than when just one of them is noisy. 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 in the number of documents retrieved. 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 text, and keywords and image is higher than the correlation between text and image.

7.2.2 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 across all five users for each query. Results are shown in Table 7.4. 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

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

Table 7.4: Precision for each query and topics representation.

is that the textual and image labels are assigned using automatic algorithms (see Section 7.1.3) which can make mistakes and assign bad labels to topics.

Queries such as *Sports*, *Health*, *Religion* and *War - Civil War* are in the top-3 precision for all 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 similar high correlation achieved between keywords and text (0.83), and keywords and image (0.84). Correlation between textual and image labels is lower (0.79) showing that there is a diversity between the queries for which the two methods achieve high/low precision. This is also likely to happen because of errors in the automatic topic labelling process.

7.2.3 Post-task

Table 7.5 shows the average scores of the answers of the participants to the 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 quite close (4.26) and image labels slightly lower (4.00). That demonstrates different topic representations can be complementary in topic browsers providing users with alternative ways to explore a document collection.

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

Finally, we observed that the participants found the retrieval task quite challenging (3.53). This might reflect the nature of the task and the limited time required to perform each query.

Question	Average
Usefulness (Keywords)	4.33
Usefulness (Text)	4.27
Usefulness (Image)	4.00
Query Familiarity	4.40
Easy of Search	3.53

Table 7.5: Results of the post-task questionnaire.

7.3 Summary

This chapter applied the methods developed earlier for labelling topics within an exploratory search task. 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 (textual and images) 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 (Lau et al., 2011) (see also Chapter 5 and 6. 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.

CONCLUSIONS

This thesis presented a variety of methods for making the output of topic models more comprehensible and useful to humans. This chapter summarises the tasks, findings and contributions presented throughout the thesis and indicates possible directions of future work.

8.1 Summary of Thesis

Chapter 2 introduced the modelling of document collections using statistical methods and the notion of topic models. We presented a variety of topic models which we later used in our experiments. In addition, we described information systems that make use of topic modelling to organise and visualise the content of unstructured large document collections and pointed out their main shortcomings. Next, we presented previous work on improving the output of topic models. We reviewed methods for computing topic coherence, labelling topics and estimating topic similarity. Finally, we described vector space models of word meaning where words are represented as vectors in high dimensional spaces where each dimension represents a context word.

Chapter 3 presented novel methods for automatically determining the coherence of topics. It proposed a novel approach where each topic word is represented as a vector in a vector space. Vector elements are weighted using either PMI or NPMI. We also experimented with different number of context terms. First, we

made use of a vector space consisting of the 5,000 most frequent words in our reference corpus. Second, we made use of two reduced spaces: using β of the most related context features given each topic word and using only the topic words of each topic. All methods are evaluated by measuring correlation with human judgements on three different sets of topics. Results obtained indicate that the measure based on topic word space outperforms previous approaches on the task.

Chapter 4 explored methods for computing semantic similarity between topics. Approaches to computing topic similarity have been described in the literature but they have been restricted to using information from the word probability distribution to compare topics and have not been directly evaluated. We addressed these limitations by providing a systematic evaluation of approaches to computing similarity between topics. We compared methods based on using distributional representations of topic words in various semantic spaces, i.e. from a reference corpus, the topic model itself and the training corpus. We also compared popular knowledge-based metrics. The chapter also introduced a data set consisting of pairs of topics together with human judgements of similarity to evaluate the proposed approaches. The data set has been made publicly available. Results demonstrated that the distributional semantic methods in the reference corpus and ESA, a state-of-the-art knowledge-based lexical similarity metric, perform better than metrics based on the comparison of the per-topic word probability distributions.

Chapter 5 introduced a novel graph-based approach to associating topics with textual labels. The proposed method takes as an input a topic and its candidate labels and the aim is to select the most appropriate one. The method makes use of topic keywords to form a query and retrieve relevant information from a search engine. A graph is generated from the words contained in the search results and these are then ranked using the PageRank algorithm. The candidate label with the highest PageRank sum of its constituent words is selected for the topic. Evaluation on a standard data set shows that the proposed method consistently outperforms the best performing previously reported supervised method, and achieves significantly better performance than the best previous reported unsupervised method.

Chapter 6 introduced the novel task of labelling topics using images. The

approach uses pictures from Wikipedia to generate a set of candidate images for each topic. The most suitable image is selected using a graph-based approach that makes use of both textual and visual information. The ranking method 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. The method is evaluated using a data set created for this study that was annotated by crowdsourcing. The data set consisting of topics and candidate images has been made publicly available. Results of the evaluation show that the proposed method significantly outperforms two baselines and the Google Image Search.

In Chapter 7 we compared a variety of topic representations within an exploratory browsing interface by applying techniques developed in Chapter 5 and 6. The representations include: (1) lists of topic keywords, (2) textual labels, and (3) image labels. Three versions of the browsing interface were created, each using one of these representations. An experiment was carried out in which users are asked to retrieve relevant documents using the interface given 20 queries on diverse subjects. Results indicated that automatically generated labels assist in representing topics within browsing 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.

8.2 Evaluation of Thesis Goals

The main aim of this thesis, as stated in the introduction, is to improve topic models by making their output more comprehensible and usable to humans. This has a direct impact on developing more efficient exploratory browsing systems for organising large volumes of text. We achieved this aim by tackling four sub-problems: (1) computing topic interpretability so that meaningless topics can be reliably identified and filtered-out; (2) identifying topics with similar themes; (3) summarising the main theme of topics using either text or images; and (4) applying topic labelling techniques to access information in document collections. The first subproblem is addressed in Chapter 3 by introducing methods which provide

more reliable estimations of topic coherence. The second subproblem is addressed in Chapter 4 by proposing more accurate topic similarity metrics than the ones previously used. Chapters 5, 6 deal with the third point by proposing novel approaches for associating topics with textual or image labels. Finally, Chapter 7 addresses the fourth subproblem by comparing different topic representations in an exploratory search task.

8.3 Future Directions

Methods proposed as part of this thesis can be extended in a number of possible ways or can be generally used in other applications. We mention some future directions:

- **Topic Coherence and Similarity**

The methods for computing topic coherence and similarity are based on vector space representations which make use of standard bag-of-words or topic model approaches. A possible way to extend these methods is by using state-of-the-art neural representations (Mikolov et al., 2013) of topic words, i.e. skip-gram vectors. Neural representations have proved to produce state-of-the-art performance in various tasks (Huang et al., 2012; Mikolov et al., 2013; Turian et al., 2010; Zhila et al., 2013) and can be used for computing topic coherence and similarity.

- **Topic Labelling**

Chapter 5 and 6 presented methods for labelling topics using textual and image labels. These labels assist in the interpretation of the topics as shown in Chapter 7. In addition to generating labels, summarisation techniques (Nenkova and McKeown, 2012) could be applied to generate a short extractive summary of each topic. Topic summaries could assist with the interpretation of topics by providing more information than keyword lists, short keyphrases or images. The summary could consist of a small number of sentences identified in the documents with the highest marginal probability given the topic. Alternatively, external sources, i.e. the Web or

Wikipedia, can be used to identify potential candidate sentences.

The method for labelling topics using images presented in Chapter 6 represents images using textual and visual features without generating a joint space. Recent studies have proposed methods for incorporating textual and visual features to representing words in joint vector spaces of multiple modalities (Bruni et al., 2011; Feng and Lapata, 2010a; Kiela et al., 2014; Lazaridou et al., 2014). These methods can be used to generate a vector space of text and visual features for each image. Image vectors can be used to compute image similarity which represents edge weights in the candidate image graph (see Section 6.1.3).

- **Topic Browsing Systems**

In Section 2.3, we identified the limitations of current topic browsers. The methods for computing topic coherence and similarity, and generating textual and image labels, described in this thesis, can be integrated into new exploratory browsing systems. The efficient post-processing of the output of topic models can provide a better browsing experience for users of such systems while making them better alternatives to standard keyword-based information retrieval systems.

BIBLIOGRAPHY

- Eneko Agirre and Aitor Soroa. 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, 2009.
- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '09)*, pages 19–27, Boulder, Colorado, 2009.
- Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, 2012.
- Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany, 2013a.
- Nikolaos Aletras and Mark Stevenson. Representing topics using images. In *Proceedings of the 2013 Conference of the North American Chapter of the Associ-*

ation for Computational Linguistics: Human Language Technologies (NAACL-HLT '13), pages 158–167, Atlanta, Georgia, 2013b.

Nikolaos Aletras and Mark Stevenson. Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers (EACL '14)*, pages 22–27, Gothenburg, Sweden, 2014a.

Nikolaos Aletras and Mark Stevenson. Labelling topics using unsupervised graph-based methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers (ACL '14)*, pages 631–636, Baltimore, Maryland, 2014b.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Representing topics labels for exploring digital libraries. In *Proceedings of the International Digital Libraries Conference (DL 2014)*, London, UK, To Appear.

Loulwah AlSumait, Daniel Barbar, James Gentle, and Carlotta Domeniconi. Topic Significance Ranking of LDA Generative Models. In Wray Buntine, Marko Grobelnik, Dunja Mladeni, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5781 of *Lecture Notes in Computer Science*, pages 67–82. Springer Berlin Heidelberg, 2009.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*, pages 25–32, Montreal, Canada, 2009.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226, 2009.

David Blei and John Lafferty. Correlated topic models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 147–154. MIT Press, Cambridge, MA, 2006.

- David M. Blei and Michael I. Jordan. Modeling Annotated Data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '03)*, pages 127–134, Toronto, Canada, 2003.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial German Society for Computational Linguistics and Language Technology Conference (GSCL '09)*, pages 31–40, Potsdam, Germany, 2009.
- Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 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, 2007.
- Elia Bruni, Giang Binh Tran, and Marco Baroni. 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, 2011.
- Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the workshop on “WordNet and other Lexical Resources” at the Second Annual Meeting of the North American Association for Computational Linguistics*, pages 29–34, Pittsburgh, PA., 2001.
- Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, Baltimore, Maryland, 2014.
- Hau Chan and Leman Akoglu. External evaluation of topic models: A graph mining approach. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 973–978, Dallas, Texas, USA, 2013.

- Allison June-Barlow Chaney and David M. Blei. Visualizing topic models. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 419–422, Dublin, Ireland, 2012.
- Jonathan Chang, Jordan Boyd-Graber, and Sean Gerrish. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '09)*, pages 288–296, Vancouver, B.C., Canada.
- Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management (CIKM '13)*, pages 209–218, San Francisco, California, USA, 2013.
- Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452, Austin, Texas, 2012.
- Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada, 1989.
- Stephen Clark. Vector space models of lexical meaning. *Handbook of Contemporary Semantics*, 2012.
- Bruce W. Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley, 2009.
- James R. Curran. *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh, 2003.
- Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2):1–60, 2008.

- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- Katrin Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012.
- Yansong Feng and Mirella Lapata. 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, 2010a.
- Yansong Feng and Mirella Lapata. 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, 2010b.
- J.R. Firth. *A Synopsis of Linguistic Theory, 1930-1955. Studies in linguistic analysis*. Blackwell Publisher, Oxford, England, 1957.
- Evgeniy Gabrilovich and Shaul Markovitch. 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, 2007.
- Debasis Ganguly, Manisha Ganguly, Johannes Leveling, and Gareth J.F. Jones. 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)*, pages 1103–1104, Dublin, Ireland, 2013.
- Matthew J Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringer, and Kevin Seppi. The Topic Browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*, Vancouver, B.C., Canada, 2010.

- Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Springer, 1994.
- Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Trans. Intell. Syst. Technol.*, 3(2):23:1–23:26, 2012.
- Karl Grieser, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg. Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness. *Journal on Computing and Cultural Heritage (JOCCH)*, 3(3):10:1–10:20, 2011.
- Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5228–5235, 2004.
- Aria Haghighi and Lucy Vanderwende. 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, 2009.
- David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii, 2008.
- Zellig Sabbettai Harris. Distributional structure. *Word*, 10:146–162, 1954.
- Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing PageRank. Technical Report 2003-35, Stanford InfoLab, 2003.
- Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, pages 957–966, Hong Kong, China, 2009.

- Alexander Hinneburg, Rico Preiss, and René Schröder. 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, 2012.
- Thomas Hofmann. 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, 1999.
- Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, 2012.
- Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 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, 2013.
- Aminul Md. Islam and Diana Inkpen. Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, pages 1033–1038, Genoa, Italy, 2006.
- Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.
- Yushi Jing and Shumeet Baluja. PageRank for product image search. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, pages 307–316, Beijing, China, 2008.
- Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In Paul Clough, Colum Foley, Cathal Gurrin, GarethJ.F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch, editors, *Advances in Information*

- Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 165–176. Springer Berlin Heidelberg, 2011.
- Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 835–841, Baltimore, Maryland, 2014.
- Dongwoo Kim and Alice Oh. Topic chains for understanding a news corpus. In *Computational Linguistics and Intelligent Text Processing*, pages 163–176. Springer, 2011.
- Solomon Kullback and Richard Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 27(1):79–86, 1951.
- Thomas K Landauer and Susan T Dumais. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.
- Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. Best topic word selection for topic labelling. In *The 23rd International Conference on Computational Linguistics (COLING ’10)*, pages 605–613, Beijing, China, 2010.
- Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 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, 2011.
- Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL ’14)*, pages 530–539, Gothenburg, Sweden, 2014.

- Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland, 2014.
- Michael Lesk. 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, 1986.
- Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pages 577–584, Pittsburgh, Pennsylvania, 2006.
- David G. Lowe. Object Recognition from Local Scale-invariant Features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157, Kerkyra, Greece, 1999.
- David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- Will Lowe. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 576–581, Edinburgh, Scotland, 2001.
- Davide Magatti, Silvia Calegari, Davide Ciucci, and Fabio Stella. Automatic Labeling of Topics. In *Proceedings of the 9th International Conference on Intelligent Systems Design and Applications (ICSDA '09)*, pages 1227–1232, Pisa, Italy, 2009.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- Xian-Li Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. Automatic labeling hierarchical topics. In *Proceedings of the 21st*

- ACM International Conference on Information and Knowledge Management (CIKM '12)*, Sheraton, Maui Hawaii, 2012.
- Qiaozhu Mei and ChengXiang Zhai. 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, 2005.
- Qiaozhu Mei, Xuehua Shen, and Cheng Xiang Zhai. 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, 2007.
- Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pages 404–411, Barcelona, Spain, 2004.
- Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- David Milne. Computing semantic relatedness using Wikipedia’s link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*, Hamilton, New Zealand, 2007.
- David Milne and Ian H. Witten. An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30, Chicago, Illinois, 2008.
- David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., 2011.

- Claudiu C. Musat, Julien Velcin, Stefan Trausan-Matu, and Marian A. RizoIU. Improving topic evaluation using conceptual knowledge. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI '11)*, pages 1866–1871, 2011.
- Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.
- David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *J. Mach. Learn. Res.*, 10:1801–1828, 2009.
- David Newman, Timothy Baldwin, Lawrence Cavedon, Eric Huang, Sarvnaz Karimi, David Martinez, Falk Scholer, and Justin Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):169–175, 2010a.
- David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '10)*, pages 100–108, Los Angeles, California, 2010b.
- David Newman, Edwin V Bonilla, and Wray Buntine. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems*, pages 496–504, Granada, Spain, 2011.
- Mark S. Nixon and Alberto S. Aguado. *Feature Extraction and Image Processing*. Academic Press, 2 edition, 2008.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.
- Alexandre Passos, Hanna M Wallach, and Andrew McCallum. Correlations and anticorrelations in LDA inference. In *Proceedings of the 2011 Workshop on*

Challenges in Learning Hierarchical Models: Transfer Learning and Optimization (held in conjunction with NIPS), 2011.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pages 248–256, Singapore, 2009.

Eduardo H. Ramirez, Ramon Brena, Davide Magatti, and Fabio Stella. Topic Model Validation. *Neurocomputing*, 76(1):125–133, 2012.

Tony Rose, Mark Stevenson, and Miles Whitehead. The reuters corpus volume 1—from yesterday’s news to tomorrow’s language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC '02)*, volume 2, pages 827–832, Las Palmas, Canary Islands, Spain.

Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1983.

Gerard Salton, Andrew Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

Koen E.A. Sande, Theo Gevers, and Cees G. M. Snoek. 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, 2008.

Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, 2014.

Alison Smith, Jason Chuang, Yuening Hu, Jordan Boyd-Graber, and Leah Findlater. Concurrent visualization of relationships between words and topics in topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 79–82, Baltimore, Maryland, USA, 2014a.

- Alison Smith, Timothy Hawes, and Meredith Myers. Hierarchy: Visualization for hierarchical topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 71–78, Baltimore, Maryland, USA, 2014b.
- Justin Snyder, Rebecca Knowles, Mark Dredze, Matthew Gormley, and Travis Wolfe. Topic models and metadata for visualizing text corpora. In *Proceedings of the 2013 NAACL-HLT Demonstration Session*, pages 5–9, Atlanta, Georgia, 2013.
- Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP '12)*, pages 952–961, Jeju Island, Korea, 2012.
- Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag Inc, 2010.
- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 384–394, Uppsala, Sweden, 2010.
- Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.
- Vladimir N Vapnik. *Statistical learning theory*. Wiley, New York, 1998.
- Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual Interna-*

tional Conference on Machine Learning (ICML '09), pages 1105–1112, Montreal, Quebec, Canada, 2009.

Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. Mining common topics from multiple asynchronous text streams. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*, pages 192–201, Barcelona, Spain, 2009.

Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 153–162, Washington, DC, 2010.

Xing Wei and W. Bruce Croft. 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, 2006.

Alisa Zhila, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomas Mikolov. Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1000–1009, Atlanta, Georgia, 2013.