MULTI-RESOLUTION MODELLING OF TOPIC RELATIONSHIPS IN SEMANTIC SPACE

Wang Wei
School of Computer Science
University of Nottingham Malaysia Campus
Jalan Broga, 43500, Semenyih, Selangor, Malaysia
Email: eyx6ww@nottingham.edu.my

Andrzej Bargiela
School of Computer Science
University of Nottingham Malaysia Campus
Jalan Broga, 43500, Semenyih, Selangor, Malaysia
Email: abb@cs.nott.ac.uk

KEYWORDS

Document modelling, Probabilistic topic models, Latent semantic allocation, Topic hierarchy

ABSTRACT

Recent techniques for document modelling provide means for transforming document representation in high dimensional word space to low dimensional semantic space. The representation with coarse resolution is often regarded as being able to capture intrinsic semantic structure of the original documents. Probabilistic topic models for document modelling attempt to search for richer representations of the structure of linguistic stimuli and as such support the process of human cognition. The topics inferred by the probabilistic topic models (latent topics) are represented as probability distributions over words. Although they are interpretable, the interpretation is not sufficiently straightforward for human understanding. Also, perhaps more importantly, relationships between the topics are difficult, if not impossible to interpret. Instead of directly operating on the latent topics, we extract topics with labels from a document collection and represent them using fictitious documents. Having trained the probabilistic topic models, we propose a method for deriving relationships (more general or more specific) between the extracted topics in the semantic space. To ensure a reasonable accuracy of modeling in a given semantic space we have conducted experiments with various dimensionality of the semantic space to identify optimal parameter settings in this context. Evaluation and comparison show that our method outperforms the existing methods for learning concept or topic relationships using same dataset.

INTRODUCTION

Document modelling has its roots in Information Retrieval (IR) and aims to provide suitable document representation to facilitate efficient processing of information for retrieval systems. One of the classic IR models, the Vector Space Model (Baeza-Yates and Ribeiro-Neto, 1999), represents a document as “bag-of-words” which is a high dimensional word vector of weighted terms that are computed by combining the “terms frequency” and “inverse document frequency” (Baeza-Yates and Ribeiro-Neto, 1999). The retrieval process is realised by computing Cosine similarity between the query and document vectors. Another classic model, Probabilistic model (we focus on the Binary Independence model) (Robertson and Jones, 1976) makes an assumption of “binary independence” between terms in a document, which is represented as a high dimensional binary vector in the word space. The retrieval process with regard to a query is done by estimating weights of each term presented in the query. The two classic and some extended IR models emphasise the co-occurrence of terms between queries and documents, for example, as the Vector Space model, “term frequency” is also included in the popular Okapi BM25 (Robertson et al., 1998) term weighting scheme.

One of the problems associated with the above-mentioned two models is that if terms do not co-occur in queries and documents, the retrieval performance deteriorates significantly. This problem is attributed to a great extent to the phenomenon of synonymy and polysemy [Deerwester1990] present in natural languages. There have been some works towards deriving low dimensionality representation of the documents in the so-called semantic space, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), probabilistic variants of Latent Semantic Analysis (pLSA) (Hofmann, 1999), and Latent Dirichlet Alllocation (LDA) (Blei et al., 2003; Steyvers and Griffiths, 2005; Griffiths and Steyvers, 2004; Steyvers et al., 2006). The intuition is that similarity between documents in the low dimensional semantic space could be higher than the one in the high dimensional word space, even if terms do not co-occur in queries and documents. Among these models, probabilistic topic models (Blei et al., 2003; Steyvers and Griffiths, 2005; Steyvers et al., 2006) aims to search for richer representations of the structure of linguistic stimuli and support the process of human cognition. The technique enables construction of a lower dimensional representation of a document while preserving its semantic structure. Furthermore, the learned topics can be interpreted as probability distribution of words, which can be understood by humans. However, the task of interpretation is not trivial. One needs to scan through the words noting corresponding probability values in learned topics and apply this insight in order to understand the meaning of a topic (i.e., associating a label to a topic). In this

Proceedings 23rd European Conference on Modelling and
Simulation ©ECMS Javier Otamendi, Andrzej Bargiela,
José Luis Montes, Luis Miguel Doncel Pedrera (Editors)
ISBN: 978-0-9553018-8-9 / ISBN: 978-0-9553018-9-6 (CD)

context, the relationships between these implicit topics are not clear (i.e., more general or more specific) and the issue has not been discussed in the literature.

In our proposed method, instead of directly operating on the latent topics, we extract topics from a document collection and make use of explicit labels that are intuitive for human understanding. The extracted topics are represented using fictitious documents, compiled as words from the vicinity of the occurrences of the topic labels, or as documents, which are annotated by these labels. These documents are fed to a set of learned probabilistic topic models to derive new representations in the low dimensional semantic space. We then propose a method to derive relationships between topics using a relationship learning algorithm based on "Information Theory Principle for Concept Relationship". Experiments with various dimensionality of the semantic space have been conducted to assess the validity of the proposed method. Empirical evaluation on a representative data set shows that the accuracy of the inferred relationships is up to 85%, which is a notable improvement compared to the results generated using the existing concept relationship learning algorithms (Sanderson and Croft, 1999; Zavitsanos et al., 2007) applied to the same dataset (the comparative study can be found in (Wei et al., 2008)).

The rest of the paper is organised as follows. We first give a brief introduction to the probabilistic topic models, emphasises on the Latent Dirichlet Allocation (Blei et al., 2003; Steyvers and Griffiths, 2005). Then we elaborate our method for learning relationships between topics, in particular, justification of the method, the "Information Theory Principle for Concept Relationship", document modelling in the semantic space with various resolution, and the relationship learning algorithm. After that, we report experimental results under different parameter settings and the comparison with results generated using other approaches. Finally we conclude the paper and discuss issues related to future research.

PROBABILISTIC TOPIC MODELS

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is introduced to alleviate some of the problems associated with classic IR techniques, i.e., synonymy and polysemy, by computing document representations in a semantic space. Probabilistic extensions of the LSA such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2005) have been developed to improve the interpretation of the results generated by the LSA. By contrast to the LSA, which explores the latent semantic space using Singular Value Decomposition, probabilistic models represent semantic properties of words and documents using latent topics interpreted as word-topic and topic-document distributions. These models have shown to be effective dimension reduction techniques (Hofmann, 1999; Blei et al., 2003; Steyvers and Griffiths, 2005).

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a modelling approach based on probabilistic inference and generalises easily to new documents. It does not suffer from the two problems associated with the probabilistic Latent Semantic Analysis (pLSA), i.e., overfitting (Blei et al., 2003), and the difficulty of generalising to new documents (Steyvers and Griffiths, 2005).

Generative Process

LDA is a generative model: each word w_i in a document is generated by sampling a topic from the topic distribution, and then sampling a word from topic-word distribution. The generative process can be represented using Equation (1) (The notations we use for LDA model follow those used in [Griffiths2004,Steyvers2005]).

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \qquad (1)$$

where $P(z_i = j)$ is the probability that jth topic is sampled for the ith word token, and $P(w_i|z_i = j)$ is the probability of sampling w_i under topic j. Let $\phi^{(j)} = P(w|z = j)$ refer to multinomial distribution over words for the topic j, and $\theta^{(d)} = P(z)$ refer to multinomial distribution over topics in the document d. The ϕ and θ are model parameters that need to be estimated.

Parameter Estimation

There are various algorithms available for estimating parameters in LDA, for example, Blei et al (Blei et al., 2003) introduced the variational inference with Expectation-Maximisation algorithm (Bilmes, 1997). In this paper, we adopt the Gibbs sampling algorithm to estimate parameters in LDA as proposed in (Steyvers and Griffiths, 2005; Griffiths and Steyvers, 2004). The idea is that instead of estimating the topic-word $p(w|z)$ and document-topic $p(z|d)$ distributions directly, one can estimate the posterior probability distribution over latent variable z given the observed data conditioned on topic assignment for all the other word tokens using Equation (2) (see (Steyvers and Griffiths, 2005; Griffiths and Steyvers, 2004)).

$$P(z_i = j|\boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \qquad (2)$$

The first term on the right side of the Equation (2) represents the probability of word w under topic j, and the second term in represents the probability of topic j in the document d. Intuitively, the assignment of a word to a topic depends not only on how likely the word is associated with a topic, but also on how dominant is the topic in a document (Steyvers and Griffiths, 2005).

The Gibbs sampling algorithm starts with random assignment of word tokens to topics. Each Gibbs sample consists of topic assignments to all of the word tokens in the corpus. Samples before the "burn-in" period are discarded due to poor estimates of the posterior probability. After the "burn-in" period, a number of Gibbs samples are preserved at regular intervals to prevent correlations between samples (Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2005). The word-topic and topic-document distribution can be obtained using Equation (3) and (4).

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(.)} + W\beta} \tag{3}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{.}^{(d)} + T\alpha} \tag{4}$$

where n is count matrix, $n_j^{(w)}$ is the number of word token w assigned to topic j, $n_j^{(d)}$ is the total number of word tokens assigned to j in the corpus, $n_j^{(d)}$ is the number of token assigned to j in document d, and $n_{.}^{(d)}$ is the number of tokens in d.

Query Folding-in

The query folding-in process involves computing low dimensional representations of either queries or previously unseen documents (Deerwester et al., 1990). The computation can be done using the same algorithm as for parameter estimation during training. We propose here that new documents are folded-in to the learned LDA models using the Gibbs sampling algorithm conditioned on the topic-word probabilities , i.e., keeping the topic-word distributions fixed, and assigning each word to LDA topics in the new document.

LEARNING TOPIC RELATIONSHIP

To avoid the process of manual labeling of documents, we propose to interpret topics through the following three steps.

- extract topics from the dataset;
- represent these topics as documents in the word space;
- fold-in the documents to the trained LDA models to obtain their representations in the low dimensional semantic space.

The first step can be realised by extracting document annotations provided by humans, or by using Information Extraction techniques (Cunningham and Bontcheva, 2005). In general, a topic is a "complex" entity which is normally well-understood by human, and is often in a form of noun-phrase, for example, "Latent Dirichlet Allocation". However, the meaning of the topic is difficult for computers to interpret. The term independence assumption apparently cannot be applied here because it would destroy the intentional structure of the entity and the meaning might be totally different from the intended one.

We represent topics using fictitious documents, which explain those topics in details and formalise a concrete context in which topics are used, for example, words in its vicinity, or documents that are annotated with the topic. The idea coincides with the "distributional hypothesis" (Harris, 1968) which states that similar words tend to appear in similar contexts. These documents are then folded-in to the learned LDA models to derive their low dimensional representations.

Information Theory Principle for Concept Relationship

Before we introduce the "Information Theory Principle for Concept Relationship", The definition for Kullback-Leibler divergence (KL divergence or $D_{KL}(P||Q)$)) (MacKay, 2003) (also known as Relative Entropy) is given to underpin the principle.

Definition 1 *The relative entropy or Kullback-Leibler divergence between two probability distributions P and Q over the same discrete space A is defined as:*
$D_{\mathrm{KL}}(P\|Q) = \sum_{i \in A} P(i) \log \frac{P(i)}{Q(i)}$

Following the Gibbs inequality MacKay (2003), KL divergence value is: $D_{KL} > 0$.

Definition 2 *A concept C_p is broader than another concept C_q if the following two conditions hold:*

- *(similarity condition) Similarity measure between C_p and C_q is greater than certain threshold TH_s (or divergence measure is less than certain threshold TH_d), and*
- *(divergence difference condition) Difference between Kullback-Leibler divergence measures: $D_{KL}(P||Q) - D_{KL}(Q||P) < 0$.*

In the above definitions, P and Q are probabilistic distributions of latent topics for concepts Cp and Cq respectively. The similarity measure can be calculated using Cosine similarity (Baeza-Yates and Ribeiro-Neto, 1999) or Jensen-Shannon (JS) divergence measures (MacKay, 2003). The threshold TH can be tuned to achieve satisfactory results.

The semantics of the KL divergence can be explained as the expected number of bits that are wasted by encoding events from the true distribution P with a code based on the not-quite-right distribution Q. It is the average "surprise" due to incoming message being drawn from distribution Q when it is expected to arrive form the true distribution P. Intuitively, given that the first condition holds, if the KL divergence $D_{KL}(P||Q)$ is less than $D_{KL}(Q||P)$, then the concept C_p is said to be more general than concept C_q . An example would help to understand the principle: if a source transmits a concept C (e.g., “Machine Learning”), and a receiver believes that the information received is the true C. However, the actual information he receives is C^* (e.g., “Neural Network”), the KL divergence $KL(C, C^*)$ measures the surprise of the receiver. If C is more general than C^*, we would expect that the first surprise $KL(C, C^*)$ will be smaller than the second one $KL(C^*, C)$.

The Definition 2 in a sense encompasses seemingly opposing tendencies. At one extreme, if two distributions are exactly the same, their respective KL divergences are same. However, calculation based on the LDA representation of the documents shows that the difference between the KL divergences of “close” or similar “distributions” can be extracted as shown in our experiments. Usefulness of the KL divergence also depends on precision of estimation of the involved distributions (which may become problematic for larger event space sizes). The experimental study shows that LDA training based on the use of Gibbs sampling is able to produce accurate estimation results as compared to other latent semantic models. Furthermore, dimensionality of the semantic space determined by the LDA model is far smaller than the original word space, thus the method does not involve computation of KL divergence in large space.

Topic Hierarchy Construction Algorithms

Interpreting relationships between topics individually does not provide an overall view on relations between topics. Based on the principle defined above, we develop a recursive algorithm for learning relationships between topics and constructing topic hierarchies, assuming that a topic can only has one “broader” topic. The basic idea of the algorithm is to search recursively for the most similar topics of the current “root” topic and remove those that do not satisfy the condition on the difference of KL divergence.

The parameters used in the algorithms are shown as follows:

- N - The total number of extracted topics.
- M_c - The maximum number of sub-nodes for a particular node.
- TH - The thresholds for similarity and divergence measures.
- TH_n - The noise factor, defined with the difference between two KL divergence measures $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$.

The parameters values of TH and TH_n can be tuned to obtain satisfactory recall, precision and $F1$ values (see Section “Evaluation”). In our experiment we have found that setting TH and TH_n within some narrow ranges results in only slight variation of recall and precision values. M_c is used to assess the effect of maximum number of sub-nodes for a particular node has on accuracy of the results (see (Wei et al., 2008) for detailed discussion on this parameter).

The algorithm starts with specifying the root topic and adding it into the Processing Vector, V. The vector V_{temp} stores the most similar topics of the current “root” node. The selected most similar concepts in V_{temp} are filtered using a two steps procedure:

- Topics whose KL divergence values with the current “root” do not satisfy the divergence difference condition are removed from V_{temp},
- For each concept left in the V_{temp}, a “broader” relation is asserted if the similarity value between the topic and the current “root” is greater than similarity value between the topic and any of the siblings of the current “root”.

The algorithm will terminate according to the conditions specified in the while loop. The pseudo-code for the algorithm is given in Algorithm 1.

```
Algorithm 1 $relationLearning(root)$
Require: Initialise $V$, $M_s$, $I$, $TH$, $TH_n$, and $M_c$.
Ensure: A terminological ontology with “broader” relations.
Initialise $V$, $M_s$, $I$, $TH$, $TH_n$, and $M_c$;
while ($i < I$ and $V$ is not empty) do
    Add current root into $V$;
    Select most similar $M_c$ nodes of root from $M_s$;
    Add similar nodes into $V_{temp}$;
    Remove nodes in $V_{temp}$ against Definition 2;
    for (all nodes $n_i$ in $V_{temp}$) do
        if (Sim($n_i$, root) > Sim($n_i$, Sibling(root))) then
            Assert broader relations between root and topic $n_i$;
        end if
        Move topic $n_i$ from $V_{temp}$ to $V$;
        Increment $i$ by 1;
    end for
    Remove current root from $V$;
end while
```

In the algorithm, the function $Sim(a, b)$ returns similarity values between nodes a and b. The function $Sibling(root)$ returns a list of siblings of the current “node”. Time complexity of the algorithm is $O(m^2 \cdot n^2)$, where $m = M_c$, $n = N$, and $n \gg M_c$. Since the value of m is much smaller than n, the algorithm is more efficient compared to the one in (Zavitsanos et al., 2007) which has time complexity of $O(n^3)$.

EXPERIMENTS

For the experiment, we have prepared a dataset consists of about 4,600 abstracts of published articles in the area of semantic Web. The dataset was processed by removing stopwords, applying Part-of-Speech (POS) tagging (only nouns, verbs, and adjectives were kept), and stemming, resulting in approximately 6900 unique words in total.

We extracted keyword annotations of the documents. Frequently appearing keywords were used as topics whose relationships were to be learned (77 topics were used in our experiment). A topic was then represented using documents annotated by the topic (i.e., documents were merged; words occurring only once were removed). The resulting document can be viewed as one that describes the particular topic. After probabilistic topic models have been trained, documents representing topics were used as "new documents" and folded into the trained topic models to obtain low dimensionality representations. This study is focused specifically on the exploration of the effect of the variable dimensionality of the semantic space and it builds on the previous publication describing in more detail the methodological issues (Wei et al., 2008).

Training Probabilistic Topic Models

We trained seven LDA models using different number of latent classes, i.e., from 30 to 90, which represents seven different latent spaces with increasing resolution of semantics. This allows to compare the results generated by the relationship learning algorithm in different semantic spaces. For each LDA model, the first 2000 Gibbs samples were dropped due to poor posterior probability estimation. After the "burn-in" period, the subsequent Gibbs samples were preserved to estimate the target probability distributions. The output of the Gibbs algorithm was a set of parameters $p(w|z)$ and $p(z|d)$. These parameters are saved into two matrices for folding-in the documents that represent the extracted topics.

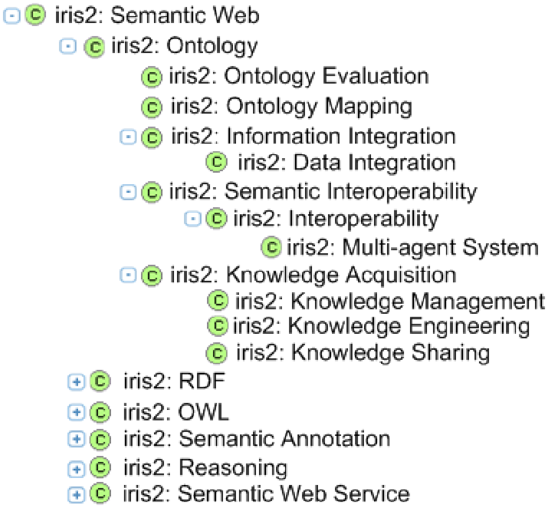
Folding-in New Documents

The topic documents were then folded-in to the seven trained LDA models respectively by using the Gibbs sampling algorithm conditioned on the topic-word probability distributions. Gibbs samples after the "burn-in" period were saved and the output was seven sets of document-topic distributions in the semantic spaces of various dimensionalities. To improve the efficiency of the relationship learning algorithm, two matrices containing pair-wise similarity and KL divergence values between the folded documents are calculated before running the algorithm.

Applying Relationship Learning Algorithm

The topic hierarchy construction procedure is a straightforward process once the topic models were learned and new documents were folded-in. In our experiment, we have used different combinations of parameters to find optimal settings, for example, the number of classes (from 30 to 90, representing various dimensionality of the semantic space), and maximum number of subnodes and the use of similarity measures. (due to the space limitation, effects of these parameters are not discussed in this paper. See (Wei et al., 2008) for more details). We have achieved a satisfying balance between the recall and precision measures by setting the range of similarity threshold $TH \subset [0.5, 0.75]$ for Cosine similarity measure, or $TH \subset [0.25, 0.45]$ for JS divergence measure, and the noise factor $TH_n \subset [0.3, 0.5]$.

A number of topic hierarchies have been constructed using the topic hierarchy learning algorithm. Figure 1 shows a snippet of a hierarchy centering on "Ontology" with "broader" relationships. The hierarchy was learned using Cosine similarity measure and LDA model with 40 classes.



- iris2: Semantic Web
  - iris2: Ontology
    - iris2: Ontology Evaluation
    - iris2: Ontology Mapping
    - iris2: Information Integration
      - iris2: Data Integration
    - iris2: Semantic Interoperability
      - iris2: Interoperability
        - iris2: Multi-agent System
    - iris2: Knowledge Acquisition
      - iris2: Knowledge Management
      - iris2: Knowledge Engineering
      - iris2: Knowledge Sharing
  - iris2: RDF
  - iris2: OWL
  - iris2: Semantic Annotation
  - iris2: Reasoning
  - iris2: Semantic Web Service

Figure 1: A snippet of the ontology centering on the topic "Ontology"

EVALUATION

The results generated by the experiments were evaluated by domain experts. Only if correctness of a relationship is agreed by all domain experts, the relationship is marked as correct. We use recall, precision and $F1$ (Baeza-Yates and Ribeiro-Neto, 1999) which are common measures in Information Retrieval to assessing performance of text retrieval, denoted as R, P, and $F1$ respectively, to evaluate the performance of our method The recall is defined as:

$$R = \frac{n_{tc}}{N_{cs}} \tag{5}$$

where n_{tc} is the number of correct learned statements given by the algorithms, N_{cs} is the total number of correct statements. It is assumed that a topic can only have one broader topic, thus the value of N_{cs} is equal to $N_c - 1$, where N_c is the total number of concepts. The precision is defined as:

$$P = \frac{n_{tc}}{N_{ls}} \tag{6}$$

where n_{tc} is the number of correct derived relationships, N_{ls} is the total number of retrieved relationships by the algorithm. Precision measure alone is not sufficient for assessing the performance of the algorithm. Low recall signifies that large portion of relationships are not learned by the algorithms. The F1 measure is defined as the harmonic mean of recall and precision.

$$F1 = \frac{2 * R * P}{R + P} \tag{7}$$

Evaluation Results

For each of the seven trained models, topic hierarchies are constructed with different maximum number of subnodes (i.e., from 5 to 10) using the algorithm. Table 1 shows the results of recall, precision and F1 under different parameters settings. Note that the numbers in the table are averaged over maximum number of sub-nodes.

Table 1: Recall measures of ontology statements based on pLSA and LDA

Settings VS NoOfClasses		**30**	**40**	**50**	**60**	**70**	**80**	**90**
Recall	COS	0.712	0.746	0.73	0.766	0.744	0.7	0.61
	JS	**0.776**	0.704	0.61	0.713	0.656	0.632	0.603
Precision	COS	0.712	**0.846**	0.732	0.771	0.785	0.737	0.722
	JS	0.8	0.76	0.709	0.837	0.743	0.736	0.768
F1	COS	0.712	**0.792**	0.731	0.768	0.764	0.718	0.661
	JS	0.788	0.731	0.656	0.77	0.697	0.68	0.675

Averaged recall (recall averaged with different number of classes and algorithm settings) of the experiment over all hierarchies is 69.3%, and the highest recall is 77.6% when "JS+LDA30" (JS divergence and 30 classes for training LDA model) is used. The highest precision with the LDA model was 84.6% when the parameters were set to 40 classes, the maximum number of sub-nodes was set to 8, and Cosine similarity measure were used. The lowest precision was about 70.9%. The best F1 value of LDA is 79.2%. We have conducted another experiment using the probabilistic Latent Semantic Analysis (pLSA) (Wei et al., 2008), the overall results were inferior than the ones generated in this experiment. We attribute the superior performance of the LDA model to its capability of generalising to new documents.

Trade-offs in Different Semantic Space

Finding optimal dimensionality of the semantic space is an issue, which echoes the considerations highlighted by "granular computing" in the general context of information-processing (Bargiela and Pedrycz, 2008, 2002). Choosing a suitable dimensionality has to take specific problems and objectives into consideration. For example, trade-offs between the emphasis on performance measures and the computational complexity have to be considered. In LDA, the computational complexity is proportional to the number of LDA classes used for training.

Evaluation of the results of the experiments with the semantic spaces of various dimensionalities allows us to develop an insight into the significance of this parameter. This is illustrated in Table 1. One of the most notable observations is that the performance measures in terms of recall, precision, and $F1$ have no strong correlations with the dimensionality of the LDA model. Results generated using LDA classes more than 60 are especially undesirable. Recall measures with dimensionalities of 30 (77.6%) and 60 (76.6%)) are higher than all the other cases, while precision measures with 40 (84.6%) and 60 (83.7%) classes are higher. The highest (averaged) F1 measure is achieved using 60 LDA classes, but the computational complexity is higher than those with lower dimensionalities. Although the choice of an appropriate dimensionality of a semantic space is determined by specific problems, we can safely conclude that the semantic space with 30, 40, and 60 classes is more appropriate for modelling topic relationships than the other alternatives considered here.

CONCLUSION AND FUTURE WORK

The recent developed probabilistic topic models, i.e., Latent Dirichlet Allocation (LDA), view the task of document modelling as a problem of probabilistic inference in which context is appropriately modulated by statistics of the environment (Steyvers et al., 2006). LDA explores a low dimension semantic space for representing documents with varying resolutions and searches for richer representation of the structures of the original documents. With careful interpretation, the learned topics are intuitively understandable for human and can be matched with the topics appearing in document collections.

Nevertheless, relationships such as "more general" or "more specific" between these topics learned using topic models are difficult to derive. In this paper, we have introduced our approach towards modelling such relationships, which does not directly operating on LDA topics. By extracting topics from a document collection and utilising LDA as effective dimension reduction technique, we are able to model relationships between topics. We have conducted experiments in the semantic space with varying resolutions to identify the optimal parameter settings for achieving satisfactory results. In addition, we have performed a comparative study with some of the existing relationship learning methods. Our method achieved notable improvement in terms of recall and precision measures. Therefore, our main contribution is the method for modelling topics relationships using probabilistic topic models, expanding the applicability of topic models to tasks other than document modelling. Another aspect of our contribution is the proposal of "Infomation Theory for Concept Relationship" which uses Kullback-Leibler divergence as a probabilistic proxy for learning topic relationships.

Although our methodology is intuitively domain independent and has produced encouraging results with a computer science publication dataset, future work will include deploying the proposed method in other domains. For datasets from highly diverse domains, one would expect that the LDA training results in several nearly disjoint sub spaces in the semantic space. Representation of a document from one domain will be dominated by LDA topics for that particular domain. However, experiments are needed to verify the statements and we would like to keep this as our future work.

REFERENCES

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Bargiela, A. and Pedrycz, W. (2002). *Granular Computing: An Introduction*. Kluwer Academic Publishers.

Bargiela, A. and Pedrycz, W. (2008). Toward a theory of granular computing for human-centered information processing. *IEEE T. Fuzzy Systems*, 16(2):320–330.

Bilmes, J. (1997). A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, ICSI-TR-97-021, University of Berkeley, California.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cunningham, H. and Bontcheva, K. (2005). Computational Language Systems, Architectures. *Encyclopedia of Language and Linguistics, 2nd Edition*.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proc Natl Acad Sci USA*, 101 Suppl 1:5228–5235.

Harris, Z. (1968). *Mathematical Structures of Language*. Wiley.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *UAI*, pages 289–296.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. pages 143–160.

Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive. In *TREC*, pages 199–210.

Sanderson, M. and Croft, W. B. (1999). Deriving concept hierarchies from text. In *SIGIR*, pages 206–213.

Steyvers, M. and Griffiths, T. (2005). Probabilistic topic models. In Landauer, T., Mcnamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Steyvers, M., Griffiths, T. L., and Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, Special issue: Probabilistic models of cognition*, 10:327–334.

Wei, W., Barnaghi, P. M., and Bargiela, A. (2008). Probabilistic Topic Models for Learning Ontology. Technical report, TR-200801, School of Computer Science, University of Nottingham Malaysia Campus.

Zavitsanos, E., Paliouras, G., Vouros, G. A., and Petridis, S. (2007). Discovering subsumption hierarchies of ontology concepts from text corpora. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 402–408, Washington, DC, USA. IEEE Computer Society.

AUTHOR BIOGRAPHIES

Wang Wei is a 3rd-year PhD student at the School of Computer Science, Faculty of Science, the University of Nottingham Malaysia Campus. He obtained his degree at the same university in 2006. His research interests include Information Retrieval, Semantic Web, Ranking, Ontology Learning, and Machine Learning. His personal webpage is at `http://baggins.nottingham.edu.my/∼eyx6ww`.

Andrzej Bargiela is Professor and Director of Computer Science at the University of Nottingham, Malaysia Campus. He is member of the Automated Scheduling and Planning research group in the School of Computer Science at the University of Nottingham. Since 1978 he has pursued research focused on processing of uncertainty in the context of modelling and simulation of various physical and engineering systems. His current research falls under the general heading of Computational Intelligence and involve mathematical modelling, information abstraction, parallel computing, artificial intelligence, fuzzy sets and neurocomputing.