A Semantic Graph based Topic Model for Question Retrieval in Community Question Answering

Long Chen
School of Computing Science
University of Glasgow
Glasgow, UK
long.chen@glasgow.ac.uk

Joemon M Jose
School of Computing Science
University of Glasgow
Glasgow, UK
joemon.jose@glasgow.ac.uk

Haitao Yu
Faculty of Library, Information and Media Science
University of Tsukuba
Tsukuba, Japan
yuhaitao@slis.tsukuba.ac.jp

Fajie Yuan
School of Computing Science
University of Glasgow
Glasgow, UK
fajie.yuan@glasgow.ac.uk

Dell Zhang
DCSIS
Birkbeck, University of London
London, UK
dell.z@ieee.org

ABSTRACT

Community Question Answering (CQA) services, such as Yahoo! Answers and WikiAnswers, have become popular with users as one of the central paradigms for satisfying users' information needs. The task of question retrieval aims to resolve one's query directly by finding the most relevant questions (together with their answers) from an archive of past questions. However, as the text of each question is short, there is usually a lexical gap between the queried question and the past questions. To alleviate this problem, we present a hybrid approach that blends several language modelling techniques for question retrieval, namely, the classic (query-likelihood) language model, the state-of-the-art translation-based language model, and our proposed semantics-based language model. The semantics of each candidate question is given by a probabilistic topic model which makes use of local and global semantic graphs for capturing the hidden interactions among entities (e.g., people, places, and concepts) in question-answer pairs. Experiments on two real-world datasets show that our approach can significantly outperform existing ones.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–Clustering; H.2.8 [**Information Systems Applications**]: Database Applications–Data mininglgorithm, Experimentation

Keywords

Community Question Answering; Question Retrieval; Knowledge Repository; Topic Modelling; Language Modelling

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

 ISBN 978-1-4503-3716-8/16/02…$15.00
DOI: http://dx.doi.org/10.1145/2835776.2835809

1. INTRODUCTION

In recent years, Community Question Answering (CQA) services, such as Yahoo! Answers, WikiAnswers, Quora, Stack Overflow, and Baidu Zhidao, have become popular knowledge sources for Internet users to access useful information online. When a user submits a new question (i.e., *query*) in CQA, the system would usually check whether similar questions have already been asked and answered before. If so, the user's query could be resolved directly by returning those archive questions (i.e., *documents*) with their corresponding answers.

Identifying relevant questions in CQA repositories is a difficult task, since the questions asked by users are always short texts which would lead to a lexical gap between the queried question and the past questions. In addition, the limited length of questions causes the sparsity of word features. To address the above limitations, researchers have proposed the use of translation models [16,33] to capture the semantic relations between words. While useful in general, the effectiveness of such models largely depends on the availability of high-quality parallel corpora (e.g., *question-answer pairs* in this context). Furthermore, simple word-level analysis model cannot handle cases where entities (e.g., *people, places,* and *concepts*) are expressed by multi-word phrases.

To go beyond the mere word-level analysis, this paper proposes a question retrieval framework on the basis of semantic graphs which makes use of an *external* resource (namely *DBpedia*) as the background knowledge to overcome the problem of the lexical gap between the queried question and the past questions. As a simple illustration, Figure 1 displays a piece of global semantic graph produced by our proposed knowledge-rich approach [1] (cf. Section 4.2.1) from two questions: "Why Kobe is better than Jordan?" and "Who is better Kobe Bryant or Lebron James?". From this graph, one can easily see that "Basketball" and "BasketballPlayer" are the central entities (i.e., *concepts* in *DBPedia*) of these two questions, even though these entities didn't appear in the original questions. Furthermore, the rich semantic connections among entities could be exploited to help inferring the

[1]https://github.com/long4glasgow/Semantics-based-Question-Retrieval

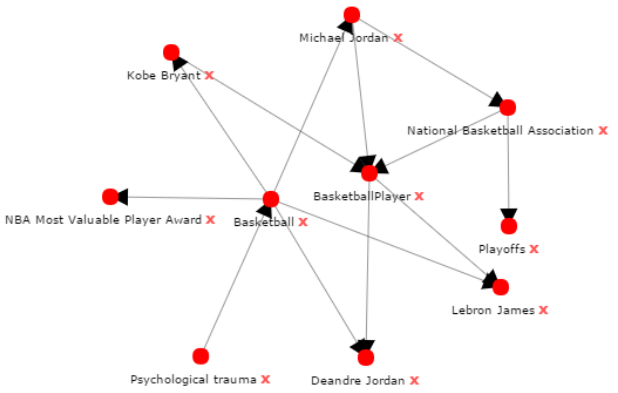


Figure 1: A piece of global semantic graph generated from two questions in Yahoo! Answers

latent topics of each question. Therefore, we would like to learn the interrelationships between entities in the global semantic graph, would allow effective information sharing among all questions in the archive. In addition to the global semantic graph, the inference of a question's latent topics could benefit from examining its similar questions according to their individual semantic graphs. In other words, if two questions have a large degree of overlap in terms of their semantic entities and relations, they probably bear a close topical resemblance to each other. Therefore, we also construct local semantic graphs for each question-answer pair in the archive with the hope to utilize their semantic similarities.

Given the semantic graphs, a novel topic modelling framework is proposed, namely, Semantic Graph based Topic Model (**SGTM**), which can seamlessly incorporate entities and relations from the semantic graphs into topic modelling. We then present a hybrid approach that blends several language modelling techniques for question retrieval, namely, the classic (query-likelihood) language model, the state-of-the-art translation-based language model, and our proposed semantics-based language model. Experiments on two real-world datasets show that the hybrid approach to question retrieval outperforms existing approaches significantly.

The rest of this paper is organized as follows. We firstly introduce the related work in Section 2. Section 3 formally defines the problem of topic modelling with semantic graph. Section 4 and 5 systematically present the proposed SGTM framework and the mixture language model. The experimental results of question clustering and question retrieval are reported in Section 6. Finally, we present our conclusion and future work in Section 7.

2. RELATED WORK

2.1 Topic Model with Network Analysis

The techniques of topic modelling, such as PLSA [12] and LDA [31], provide an elegant mathematical way to analyze large volumes of unlabelled text. Recently, a large number of studies, such as Author-Topic Model (ATM) [28] and Contextual Focused Topic Model (CFTM) [9], try to integrate some side information of network structure with topic modelling, but they mostly focus on homogeneous networks rather than heterogeneous networks. Entity-Topic Model (ETM) [18] combined LDA with entity-document relations, which is somewhat similar to our idea. However it assumes that an edge (entity-document)created in exactly the same way as a word, whereas our approach directly takes into account several types of relations (entity-document and entity-entity relations) through regularized propagation.

In Mei's seminal work [24], a homogeneous network was employed as a biased regularizer to overcome the overfitting problem of topic modelling. Deng et al. [10] later on combined Probabilistic Latent Semantic Analysis (PLSA) [12] (see Section 4.1) with the regularizer learned from a heterogeneous network. However, it was originally designed for academic networks, and thus didn't utilize the contextual information from any knowledge repositories. In addition, their framework only incorporates the heterogeneous network (i.e., relations among entities), while the homogeneous network (i.e., relations between entity pairs with weight) is completely ignored, whereas we consider both of them in our proposed framework.

2.2 Knowledge Rich Approaches

The recent advances in knowledge-rich approaches (e.g., DBPedia[2] and Knowledge Graph[3]) provide new techniques to gain insight into the semantic structure of a question archive. While enormous success has been made in several NLP tasks such as document similarity [26], topic labelling [15], and question answering [5], the feasibility and effectiveness of such knowledge-rich approaches in topic modelling and question retrieval are mostly unknown. Hulpus et al. [15] reported a framework that extracts sub-graphs from DBpedia to label the topics obtained from a topic model. However, they consider topic model and graph labelling as two separate processes, which may result in the loss of rich semantics. On the contrary, our framework discovers the latent topics and semantic network simultaneously, reinforcing the topic model performance with multi-typed relations. In addition, their graph construction process relies on a small set of manually selected DBpedia relations, which does not scale and needs to be tuned each time given a different knowledge repository. Instead, we pruned our semantic graphs by filtering and weighting the edges (see Section 4.2.2). This may look similar to [26], but their work attempts to produce graph-representation of documents for the task of document ranking, while we aim to construct semantic graphs for the task of topic modelling and question retrieval.

2.3 Question Retrieval

The technique of language modelling has proved to be effective for question retrieval in CQA. The state-of-the-art approach utilized the classic (query-likelihood) language model [34] (see Section 5.1) together with the translation-based language model [16,33] (see Section 5.2). Ji et al. [17] have recently reported an approach to question retrieval based on the question-answer topic model which uses LDA to learn the latent topics underlying the surface text of question-answer pairs. However, the semantic level of their work is substantially different from ours: their topic model is designed at word level, whereas ours relies on the external knowledge repository at the concept level, and thereby the topics in our model are more general and expressive.

[2]http://wiki.dbpedia.org/
[3]https://developers.google.com/freebase/

Cao et al. have proposed a category-based language model [6, 7] for question retrieval, but it requires users to manually assign one topic category to each of their query questions, which is not always feasible. Instead of using the manually labelled categories, Zhou et al. [36] later on improved Cao's framework using category-based entries extracted from Wikipedia, but their category-based model cannot be applied straightforwardly to incorporate topics into question retrieval, because it assumes that a question belongs to one and only one topic, while we believe that a question can have multiple topics (each to a certain degree).

3. PRELIMINARIES

In this section, we formally introduce several concepts and notations.

Definition 1 (Question): A ***question*** d in a archive D is a sequence of words $w_1w2....w_{|d|}$, where w_i is a word from a fixed vocabulary. Following a common simplification in most work in information retrieval [12], we consider each question (i.e., document) as a bag of words, and use $n(d, w)$ to denote the number of occurrence of word w in d.

Definition 2 (Entity): An entity e in our system, given that we are using DBPedia resource URIs, can be either an instance or a concept in DBPedia. The former are concrete entries of DBPedia (e.g. $dbpedia : Barrack\_Obama$), while the latter are the classes found within the DBPedia Ontology (i.e., the types of instances such as people, places, organizations).

Definition 3 (Semantic Graph): A semantic graph G consists of V the set of entities in the text data and E the set of edges representing the relations between entities. For instance, an edge $< u, v >$ is a binary directed relation from entity u to entity v, where we use $w(u, v)$ to denote the weight of $< u, v >$.

We call the semantic graph built from the entire corpus (the archive of past questions) the *global* semantic graph, and those built from individual documents (i.e., questions) *local* semantic graphs

Now we can formulate our problem of SGTM as follows. Given a question archive D and its corresponding semantic graph $G = (V, E)$, we would like to extract K major topics $Z = \{z_1, z_2, \ldots, z_K\}$ from D, where z_k is a probability distribution of words, and the probability of a word w is denoted as $P(w|z)$. What SGTM essentially does is to group similar or related questions into semantic clusters according to not only their textual similarity but also the underlying semantic graphs.

4. TOPIC MODELS

In this section, we propose a propagation algorithm to combine semantic graphs with the textual information for topic modelling, namely **Regularized Propagation**. The goal of this algorithm is to estimate the probabilities of topics for documents as well as other associated entities, in order to improve the performance of topic modelling.

4.1 Probabilistic Topic Models

In Probabilistic Latent Semantic Analysis (PLSA) [12], an unobserved topic variable $z_k \in \{z_1, ..., z_K\}$ is inferred from the occurrences of different words $w_j \in \{w_1, ..., w_M\}$ in a particular document $d_i \in \{d_1, ..., d_N\}$. The joint probability of an observed pair (d, w) can be expressed as

$$P(d_i, w_j) = P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad (1)$$

where $P(w_j|z_k)$ is the probability of word w_j occurring in topic z_k, and $P(z_k|d_j)$ is the probability of topic z_k for document d_i. The model parameters can be estimated by maximizing the log likelihood of the document collection D

$$L(D) = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad (2)$$

through Expectation-Maximization (EM) [12].

PLSA provides a simplistic solution to find topics of documents. There is no guarantee that documents are associated with similar topics. Furthermore, in PLSA there is no constraint on the parameters $\theta_{ki} = P(z_k|d_i)$, the number of which grows linearly with the data. Therefore, the model tends to overfit the data. Latent Dirichlet Allocation (LDA) [31] which is essentially the Bayesian version of PLSA, can alleviate the overfitting problem, but it still ignores the semantic relationships among documents. In this paper, we propose to use the semantic graphs of the document collection to enhance PLSA topic modelling.

4.2 Semantic Graph based Topic Model

Generally speaking, an entity $e \in V$ is likely to be relevant to a specific topic of a question if its adjacent entities in the semantic graph are largely relevant to that topic. Following this intuition, we define a regularization term for the learning of topic model:

$$R_V(G) = (1-\mu) \sum_{i=1}^{|D|} \sum_{k=1}^{K} (P(z_k|d_i) - \sum_{e_i \in V_l} \sum_{e_u \in V_{d_i}} P(z_k|e_i)w(e_i|e_u))^2 + \mu \sum_{i=1}^{|D|} \sum_{k=1}^{K} (P(z_k|d_i) - \sum_{e_j \in V_g} \sum_{e_u \in V_{d_i}} P(z_k|e_j)w(e_j|e_u))^2 \quad (3)$$

where μ is a bias parameter which strikes the optimal balance between the local semantic graph and the global semantic graph. V_l and V_g are two set of entities derived by local semantic graph and global semantic graph (see Section 4.2.1), respectively. $P(z_k|e_i)$ and $P(z_k|e_j)$ are estimated in the same way as $P(z_k|d_i)$ by using the EM algorithm (see Section 5.4.1). $w(e_i|e_u)$ is the weight between entity e_i and e_u in local semantic graph, $w(e_j|e_u)$ is the weight between entity e_j and e_u in global semantic graph (see Section 4.2.2).

To incorporate both the textual information and the semantic graph into the topic model, we define a **regularized propagation** framework by adding the regularization term with weight λ to the log-likelihood as follows:

$$L'_{rp}(D) = -(1-\lambda)L(D) + \lambda R_V(G_D) \quad (4)$$

where R_V is defined in equation (3). By minimizing R_V, we will smooth the topic distribution on the semantic graphs, where adjacent entities would have similar topic distributions. On the other hand, by minimizing $-L(C)$, we will find $P(z_i|d)$ and $P(w|z_i)$ which fit the text as much as possible. The parameter $\lambda \in [0, 1]$ is set to control the balance between the log-likelihood and the smoothness of topic distributions over the semantic graphs. When $\lambda = 0$, the

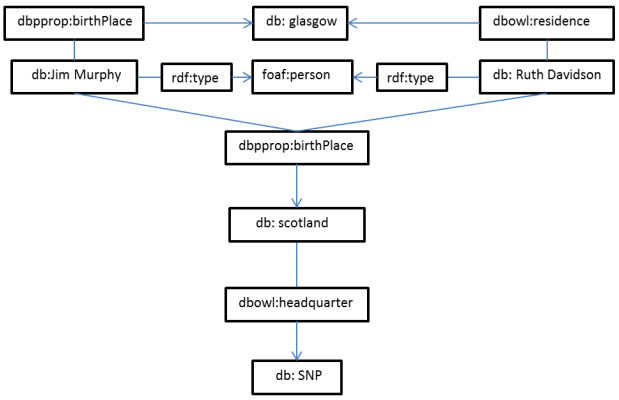


Figure 2: A sample semantic graph

objective function backs off to the standard PLSA. Minimizing $L'_{rp}(D)$ would only give us the topics which best fit the content of the document collection D. When $\lambda = 1$, the objective function backs off to R_V, which can be deemed as a "loss function" to measure how well the topic distributions on the semantic graph are consistent with the topic distribution on the documents. This is related to the objective of spectral clustering (e.g., ratio cut [8]). By minimizing $L'_{rp}(D)$, we can extract question clusters that making use of not only the text content of documents but also the structure of semantic graphs.

4.2.1 Mapping Questions into Semantic Graphs

When computing $P(e_j|e_u)$ in the above SGTM model, the method of [26] is adopted to construct the semantic graph. We start with a set of input entities C, which is found by using the off-the-shelf entity recognition tool DBpedia Spotlight[4].

We then create a directed graph G as follows: 1) we define the set of entities V of G to be made up of all input entities, i.e., we set $V := C$; 2) we connect the entities in V based on the directed paths found between them in DBpedia. Specifically, the set of entities in V are expanded into a graph by conducting a depth-first search along the DBpedia graph and adding all the visited relations and entities, to a certain limit. So the finally constructed semantic graph consists of all the "seed" entities identified from the documents together with all the edges found along the paths up to maximal length L that connect them. In this work, we set $L = 2$, as we find that the model with $L > 2$ tends to produce very large graphs and introduce lots of noise.

Figure 2 illustrates an example of a semantic graph generated from the set of entities { **db:Jim Mruphy, db:Ruth Davidson, db:SNP** }, which are found in the question "Have Jim Murphy and Ruth Davidson won SNP's general election?" Starting from these seed entities, we conduct a depth-first search to add relevant intermediate entities and relations to G (e.g., **db:scotland** and **foaf:person**). As a result, we obtain a semantic graph with additional entities and edges which provide us with rich knowledge about the original entities. Please note that we create two kinds of semantic graphs, namely, local semantic graph and global semantic graph. A local semantic graph is built for an individual question (associated with its answers) in order to detect its pairwise contextual similarities with other questions. The global semantic graph is constructed from the entities in the entire question archive, in order to capture the global contextual information of all existing questions.

[4]https://github.com/dbpedia-spotlight/dbpedia-spotlight

4.2.2 Semantic Relation Weighting

So far, we simply traverse a set of input entities from DBpedia graph. However, DBpedia ontology contains semantic relations at different levels which may not be equally informative. For example, in Figure 2, the seed entities **db:Jim Murphy** and **db:Ruth Davidson** can be connected through both **rdf:type foaf:person** and **dbpprop:birthPlace**, but the former is less informative since it can apply to a large number of entities (i.e., all people in DBpedia). We can use real-valued weights to describe the degree of correlation between entities in the graph, and the core idea underlying our weighting scheme is to reward those edges that are most specific to the entities connected by them. Therefore we define the weighting function as

$$W = -\log(P(W_{Pred})) \tag{5}$$

where W is the weight of an edge, $P(W_{Pred})$ is the probability that the predicate W_{Pred} (such as **rdf:type**) is describing the specific semantic relation. This measure is based on the hypothesis that specificity is a good estimator for relevance. We can compute the document frequency for each type of predicates, as we have the whole DBpedia database available and are able to query for all possible realizations of the variable X_{Pred}. $P(W_{Pred})$ is then defined in a similar way as the tf-idf [23] representation of W_{Pred}. In our example, an edge labelled with **rdf:type** will accordingly get a weight W which is considerably lower than those labelled with **dbpprop:birthplace**.

There are often multiple relations between two entities, so the relation with the highest weight will be selected as the final edge. For instance, in the above example, **db:Jim Murphy** and **db:Ruth Davidson** can be connected by **db:glasgow**, **foaf:person**, and **db:scotland**. The path to **db:glasgow** will be selected since **dbowl:residence** has a higher weight than **dbpprop:birthPlace** and **rdf:type**.

5. RETRIEVAL MODELS

5.1 Classic Language Model

Using the classic (query-likelihood) language model [34] for information retrieval, we can measure the relevance of an archive question d with respect to the query question q as:

$$P_{cla}(q|d) = \prod_{w \in q} P_{cla}(w|d) \tag{6}$$

assuming that each term w in the query q is generated independently by the unigram model of document d. The probabilities $P_{cla}(w|d)$ are estimated from the bag of words in document d with Dirichlet prior smoothing.

5.2 Translation-based Language Model

There are often lexical gaps between a query question and archive questions in CQA. For example, "Where can I see movies for free online" and "Anyone share me a DVD streaming link?" probably have the same meaning but are

expressed in quite different words. It has been demonstrated that this issue could be addressed by the translation-based language model [16, 33]:

$$P_{tra}(q|d) = \prod_{w \in q} P_{tra}(w|d) \quad (7)$$

$$P_{tra}(w|d) = \sum_{t \in d} P(w|t)P(t|d) \quad (8)$$

where $P(w|t)$ represents the probability of a document term t being translated into a query term w. As in [33], we estimate such word-to-word translation probabilities $P(w|t)$ on a parallel corpus that consists of 200,000 archived question-answer pairs from Yahoo! Answers.

5.3 Topic-based Language Model

There could be different topics underlying different questions. In this paper, we propose to take the latent topics into account for question retrieval in the language modelling framework:

$$P_{top}(q|d) = \prod_{w \in q} P_{top}(w|d) \quad (9)$$

$$P_{top}(w|d) = \sum_{k=1}^{N} P(w|z_k)P(z_k|d) \quad (10)$$

where z_k represents a latent topics learned from the TMBP model [10], $P(w|z_k)$ is the unigram language model of topic z_k, and $P(z_k|d)$ is the probability that document d belongs to topic z_k. In this model, topics are captured at the concept level since an entity network is incorporated into the topic modelling framework. However, instead of exploiting external knowledge, such as DBpedia, the topic distribution of each entity is learned indirectly from its corresponding documents.

5.4 Semantics-based Language Model

The semantics-based language model proposed by us aims to capture the latent topics in a better way via exploiting the hidden interactions among different entities in the SGTM. It can be descried formally as follows:

$$P_{sem}(q|d) = \prod_{w \in q} P_{sem}(w|d) \quad (11)$$

$$P_{sem}(w|d) = \sum_{k'=1}^{N} P(w|z'_k)P_E(z'_k|d) \quad (12)$$

where z'_k represents a latent topics, $P(w|z'_k)$ is its corresponding unigram language model and $P_E(z'_k|d)$ is the probability that the question d belongs to that topic (cf. Section 5.4.1).

Compared to the topic-based language model of [17], the semantics-based language model presented above is more general and more robust, because instead of learning topics solely from the document collection at the word-level, it utilizes the SGTM topics which are at multiple semantic levels, namely words, entities, and the entity relations extracted from the knowledge repository, DBpedia.

5.4.1 Model Fitting with the EM Algorithm

To estimate $P(w|z_k)$ and $P(z_k|d)$ in the above model, we use the Expectation Maximization (EM) algorithm, which alternates two steps, E-step and M-step. The unobserved latent variables in our model include $\phi = P(w_j|z_k)$, $\theta = P(z_k|d_i)$, and $\varphi = P(z_k|e_l)$.

Let us first consider the parameter estimation in PLSA. In E-step, we calculate the posterior probabilities $P(z_k|d_i, w_j, e_l)$:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k'=1}^{K} P(w_j|z_{k'})P(z_{k'}|d_i)} \quad (13)$$

$$P(z_k|d_i, e_l) = \frac{P(z_k|e_l)P(e_l|d_i)}{\sum_{k'=1}^{K} P(z_{k'}|e_l)P(e_l|d_i)} \quad (14)$$

In the M-step, we maximize the expected complete data log-likelihood for PLSA:

$$Q_D = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} P(z_k|d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad (15)$$

There is a closed-form solution [10] to maximize Q_D, which are listed in Equation 16, 17, and 18.

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j'=1}^{M}\sum_{i=1}^{N} n(d_i, w_{j'})P(z_k|d_i, w_{j'})} \quad (16)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j'=1}^{M} n(d_i, w_{j'})} \quad (17)$$

$$P(z_k|e_l) = \frac{\sum_{s=1}^{S} n(e_l, w_s)P(z_k|d_i, e_l)}{\sum_{s'=1}^{S} n(e_l, w_{s'})} \quad (18)$$

Thus the model parameters in PLSA could be estimated efficiently using the standard EM algorithm. However, due to the introduction of the regularizer $R_V(G)$, there is no such closed form solutions for parameter estimation in SGTM (Equation 4). Fortunately, we can use the generalized EM algorithm [25] to maximize the log-likelihood of $L'_{rp}(D)$. In the following, we will explain the model fitting procedure for $L'_{rp}(D)$.

It is easy to see that $L'_{rp}(D)$ and $L(D)$ share the same hidden variables z_k, and therefore could have the same E-step. Since the regularization $R_V(G)$ doesn't involve the parameter $P(w_j|z_k)$, we can still use the same M-step estimation for $P(w_j|z_k)$ as in Equation 16. Now the problem is how to estimate the parameter values $\phi = P(z_k|d_i)$ and $\theta = P(z_k|e_l)$. Instead of maximizing $P(z_k|d_i)$ and $P(z_k|e_l)$ directly, the generalized EM algorithm tires to improve the expected $P(z_k|d_i)$ as follows. First, we find $\theta^{(1)}_{t+1}$ using Equation 13 and 14, which maximizes Q_D instead of $L'_{rp}(D)$. Then, we start from $\theta^{(1)}_{t+1}$ and try to minimize $R_V(G)$, which can be done through the Newton-Raphson method [4]. Given a function $f(x)$ and the initial value x_t, the Newton-Raphson updating formula to decrease $f(x)$ is $x_{t+1} = x_t - \xi\frac{f'(x)}{f''(x)}$, where $0 \le \xi \le 1$ is the step parameter. With $\theta^{(1)}_{t+1}$, we can decrease $R(G)$ by updating $P(z_k|d_i)$ in each step.

$$P_E(z_k|d)^{(n+1)}_{t+1} = \xi P(z_k|d)^{(n)}_{t+1} + (1-\xi)\sum_{e \in V_d} \frac{P(z_k|e)}{|V_d|} \quad (19)$$

where $P(z_k|d_i, e_l)$ is obtained from the initial E-step, the step parameter ξ can be interpreted as a controlling factor

of smoothing the topic distribution among the adjacent entities. It repeatedly updates $P_E(z_k|d)^n_{t+1}$ until $P_E(z_k|d)^{n+1}_{t+1} \leq P_E(z_k|d)^n_{t+1}$. We summarize the generalised EM algorithm for parameter estimation in this regularized propagation framework by using generalized EM algorithm in Algorithm 1.

5.5 Mixture Model

Since the classic language model, the translation-based language model and the semantics-based language model cover different grained semantic levels, it would be beneficial to combine their strengths for question retrieval. So we can mix the above language models via linear combination:

$$P_{mix}(q|d)=\alpha P_{cla}(q|d) + \beta P_{tra}(q|d) + \gamma P_{sem}(q|d) \quad (20)$$

where α, β, and γ are three non-negative weight parameters satisfying $\alpha + \beta + \gamma = 1$. When $\gamma = 0$, the complete mixture model backs off to the current state-of-the-art approach, i.e., the combination of the classic language model and the translation-based language model [33].

Algorithm 1: Model fitting for regularized propagation

Input : Input data, which includes: $G = (V, E)$ with word occurrences $n(d_i, w_j)$. The number of topics K, Newton step parameter γ, regularization parameter λ.

Output: $\phi = P(w_j|z_k)$, $\theta = P(z_k|d_i)$, and $\varphi = P(z_k|e_l)$

1: Random initialize the probability distribution ϕ_0 and θ_0
2: $t \leftarrow 0$;
while $t < MaxIteration$ **do**
  E-step: Calculate $P(z_k|d_i, w_j)$ and $P(z_k|d_i, e_l)$ as in Eq. 13 and 14
  M-step:
  Re-estimate $P(w_j|z_k)$ as in Eq. 16
  Re-estimate $P(z_k|d_i)$ as in Eq. 17
  Re-estimate $P(z_k|e_l)$ as in Eq. 18
  $P(z_k|d_i)^1_{t+1} \leftarrow P(z_k|d_i)_{t+1}$
  Calculate $P(z_k|d_i)^2_{t+1}$ as in Eq. 19
  while $L'_{rp}(C)^2_{t+1} > L'_{rp}(C)^1_{t+1}$ **do**
    $P(z_k|d_i)^1_{t+1} \leftarrow P(z_k|d_i)^2_{t+1}$
    Calculate θ^2_{t+1}, update $P(z_k|e)_{t+1}$
  end
  if $L'_{rp}(C)(\theta^{(1)}_{t+1}) \geq L'_{rp}(C)(\theta_t)$ **then**
    $P(z_k|d_i)_{t+1} \leftarrow P(z_k|d_i)^1_{t+1}$
    update $P(z_k|e)_{t+1}$
  end
  else
    Keep current θ, ϕ
  end
  $t \leftarrow t + 1$
end

6. EXPERIMENTS

6.1 Experimental Setup

We conducted experiments on two real-world CQA datasets. The first dataset, YA, comes from Yahoo! Answers. It is part of Yahoo! Labs' Webscope[5] L6 dataset that consists of 4,483,032 questions with their answers from 1 January 2006 to 1 January 2007. The second dataset, WA, comes from WikiAnswers. It contains 824,320 questions with their answers collected from WikiAnswers[6] from 1 January 2012 to 1 June 2012.

Table 1: Statistics of the YA and WA datasets

	YA	WA
# of questions	40,000	40,000
# of entities (local)	123,263	83,541
# of entities (global)	27,324	24,750
# of relations (local)	142,454	159,492
# of relations (global)	71,719	69,713

In order to produce the semantic graphs, we first apply DBpedia Spotlight on each question-answer pair of these two datasets. We use the *subject* field as question part and the *bestanswer* field as the answer part. The text of questions and answers have been preprocessed by case-folding and stopword-removal (using a standard list of 418 common words). Given the disambiguated entities (see Section 4.2.1), we create local and global entity collections, respectively, for constructing local and global semantic graphs. The creation process of entity collections is organized as a pipeline of filtering operations:

1. The isolated entities, which have no connections with the other members of the entity collection in the DBpedia repository, would be removed, since are almost useless in the topic propagation process.
2. The infrequent entities, which appear in less than five documents when constructing the global entity collection, would be discarded.
3. Similar to the previous step, we discard entities that appear less than twice in the question archive when constructing the local entity collections.

6.2 Experiments with Topic Modelling

We first experimented with topic modelling on 40,000 questions that are randomly sampled from the top ten categories of YA and WA dataset, respectively. The statistics of these two datasets along with their corresponding entities and relations are shown in Table 1. The top ten categories distributions in these two datasets are shown in Figure 3.

We randomly split each of the dataset into a training set, a validation set, and a test set with the ratio 2:1:1. We learned the parameters of the Semantic Graph based Topic Model (SGTM) as well as several other representative topic models from the training set, tuned the parameters of each model on the validation set, and evaluated the performance of each model on the test set. To demonstrate the effectiveness of our SGTM method, we compare it with the following topic modelling techniques:

- **PLSA:** The baseline approach which only employs the classic Probabilistic Latent Semantic Analysis [12].
- **ATM:** Author Topic Model, which combines LDA with authorship network [28]. In our experiments, authors are replaced with entities.

[5]http://webscope.sandbox.yahoo.com/

[6]http://wiki.answers.com/

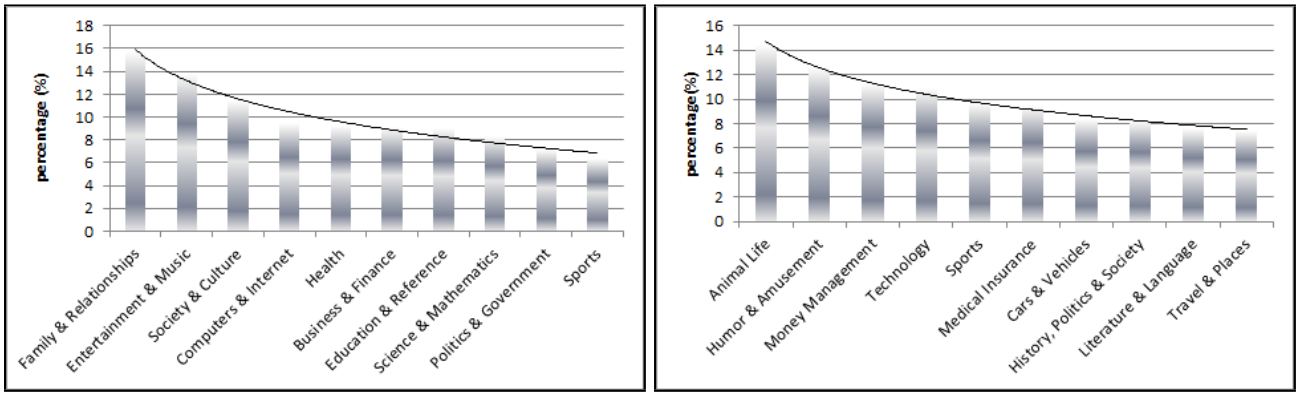


Figure 3: (a) and (b) show the category distribution of YA and WA datasets respectively

- **TMBP:** The state-of-the-art approach, Topic Model with Biased Propagation [10], which combines PLSA with an entity network (without using any external knowledge, such as DBpedia).
- **SGTM:** Our proposed Semantic Graph based Topic Model (See Section 4.2).

In order to evaluate our proposed topic model and compare it to existing ones, we use two metrics, accuracy (AC) and normalized mutual information (NMI), which are popular for evaluating the effectiveness of clustering methods. The accuracy is defined as $AC = \frac{\sum_1^n \delta(a_i, map(l_i))}{n}$ [32], where n denotes the total number of questions, $\delta(x, y)$ is the delta function that equals one if $x = y$ and zero otherwise, and $map(l_i)$ is the mapping function that maps each cluster label l_i to the corresponding label from the data corpus. Given two sets of document clusters, C and C', their mutual information is defined as: $MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}$ [32], where $p(c_i)$ and $p(c'_j)$ are the probabilities that a randomly chosen document belongs to the clusters c_i and c'_j, respectively, and $p(c_i, c'_j)$ is the joint probability that a randomly chosen document belongs to the cluster c_i and c'_j at the same time.

Parameter Setting: For PLSA, we only use question-answer pairs for question clustering with no additional entity information. For ATM, we use symmetric Dirichlet priors in the LDA estimation with $\alpha = 50/K$ and $\beta = 0.01$, which are common settings in the literature. For TMBP, an entity-based heterogeneous network is constructed, and its parameter settings were set to be identical to those in [10].

Consistent to our previous setting of top 10 categories, we set the number of topics (K) to be 10 for both YA and WA. Figure 4 shows how the SGTM clustering performance varies with the different parameter values. The essential parameters in the SGTM framework are λ and μ. As mentioned in Section 4.2, λ controls the relative importance of the inherent textual information against the semantic graph information, and μ controls the balance between the local semantic graph and the global semantic graph. When $\lambda = 0$, it is the baseline PLSA model. When $\lambda = 1$, is is entirely determined by the graph regulation term R_V (cf. Equation 3). The performance of question clustering were tuned on the validation set and evaluated on the training set through 5-fold validation. The results reported in Figure 4 are those averaged over the five trials. It can be seen that SGTM with global semantic graphs generally performs better than SGTM with local semantic graphs, which possibly suggests that the global context is more important than the local context for the purpose of question clustering. Furthermore, the best performance is achieved when combining these two with the parameter setting: $\lambda = 0.4$ and $\mu = 0.5$.

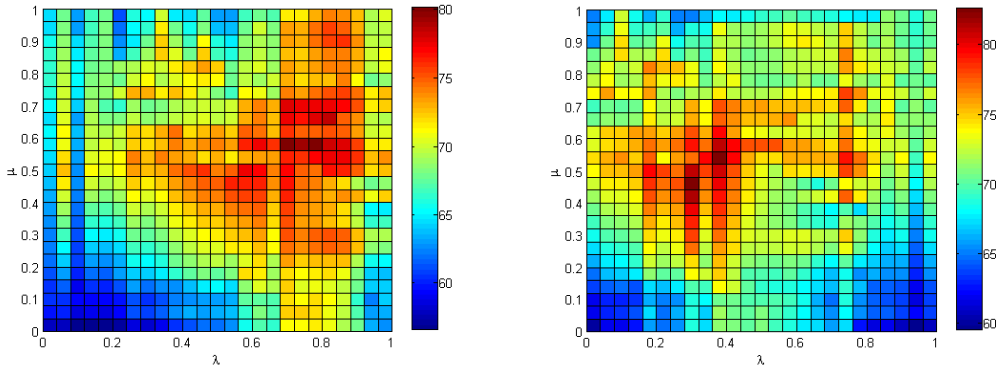
Table 2 depicts the question clustering performances of different topic modelling methods. For each method, 20 test runs are conducted on the test set, and the final performance scores were calculated by averaging the scores from those runs. It can be seen that ATM outperforms the baseline PLSA with additional entity network information. As expected, TMBP outperforms ATM since it directly incorporates the heterogeneous network of entities. Our proposed SGTM improves accuracy by 9.1% over the classical PLSA baseline, and 2.9% over the state-of-the-art TMBP on YA dataset. A comparison using the paired t-test clearly shows that SGTM outperforms all the other methods PLSA, ATM, and TMBP significantly. This indicates that exploiting the semantic graph knowledge greatly improve the performance of topic modelling.

6.3 Experiments with Question Retrieval

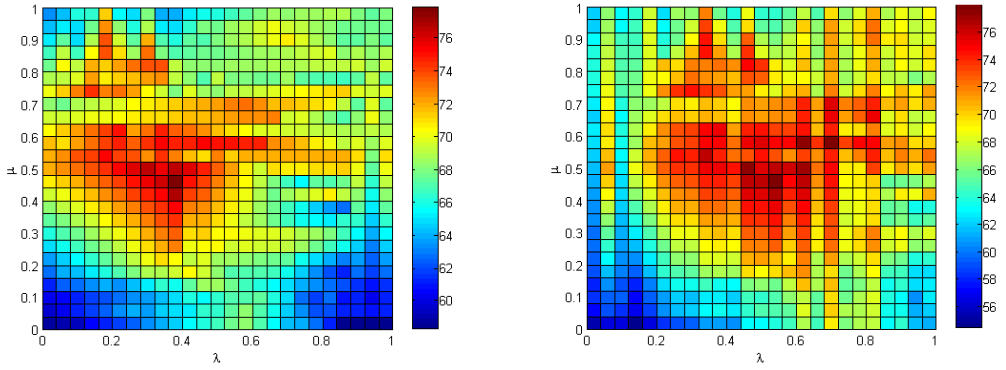
We then experimented with question retrieval with a similar setup in [33]: 50 questions were randomly sampled from the YA and WA datasets respectively for testing, and the top archive questions (i.e., search results) returned for each test query question were manually labelled as either relevant or not.

In our question retrieval experiments, we compare the following four approaches:

- the baseline approach which only employs the classic language model (C);
- the state-of-the-art approach which combines the classic language model and the translation-based language model (C+T) [33];
- the proposed hybrid approach which blends the classic language model, the translation-based language model, and the topic-based language model (C+T+T) [10];
- a hybrid approach which combines the classic language model, the translation-based language model, and the semantics-based language model (C+T+S).



(a) YA (b) WA



(c) YA (d) WA

Figure 4: (a) and (b) show the accuracy (%) of SGTM framework with varying parameters λ and μ; (c) and (d) show the NMI (%) of SGTM framework with varying parameters λ and μ.

Table 2: The clustering performance of different methods on (a) YA and (b) WA datasets (-*-* and -* indicate the statistical significance of performance decrease from that of SGTM with p-value <0.01 and p-value < 0.05, respectively).

(a) YA

	PLSA	**ATM**	**TMBP**	**SGTM**
AC	0.662-*-*	0.685-*	0.724-*	0.753
NMI	0.657-*-*	0.732-*	0.765-*	0.819

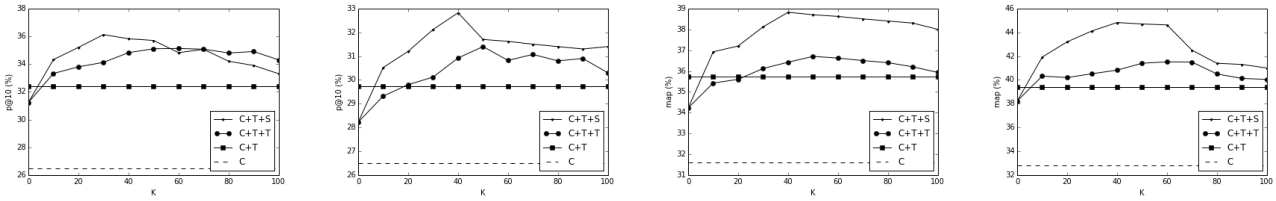
(b) WA

	PLSA	**ATM**	**TMBP**	**SGTM**
AC	0.636-*-*	0.649-*-*	0.652-*	0.694
NMI	0.654-*-*	0.689-*-*	0.717-*	0.734

Parameter Setting: All parameter values of these approaches to question retrieval were tuned according to Precision at 10 (P@10) [22] or Mean Average Precision (MAP) [22]. The mixture coefficients in Equation 20 were tuned on the training data to achieve optimal results, as shown in Table 3. In the mixture models (C+T) and (C+T+S), the ratio between parameter values α and β was set as same as those in [33]. All the other parameters were set to their optimal values that have been found in Section 6.2.

Previously the number of topics K were set a priori as 10 (the number of categories) as we need the category label for calculating the accuracy. However, using only 10 topics is not necessarily the optimal value for the question retrieval task. Furthermore, it is well known that in general we need to use more topics for larger datasets to achieve the best topic modelling effect. Hence, we tried the mixture language model with different values of K. As shown in Figure 5, it is clear that the semantics-based approach (C+T+S) achieves the best result when $K = 40$, and the topic-based approach (C+T+T) achieves the optimal result when $K = 50$. These numbers are much smaller than the corresponding optimal K value (200) reported in the question-answer topic model [17]. As we explained in Section 5.4, in the topic-based language model, the topics were solely learned from the question-answer pairs at word-level, whereas in the semantics-based language model model, the topics are obtained from the question-answer pairs and also the external knowledge repository at both the word-level and the concept-level. Such *semantic* topics for each question are more powerful and expressive in SGTM, and thereby a much smaller number of topics is needed.

Given the optimal parameter values, the retrieval performances of those approaches on the test set, measured by $P@10$ and MAP, are reported in Table 4. Consistent to the observation in [33], adding the translation-based language model (C+T) brings substantial performance improvement



(a) YA (b) WA (c) YA (d) WA

Figure 5: (a) and (b) show the P@10 (%) of retrieval models with varying number of topics K; (c) and (d) show the NMI (%) of retrieval models with varying number of topics K.

Table 3: The mixture coefficients for different question retrieval approaches.

	C	C+T	C+T+T	C+T+S
α	1	0.3	0.18	0.18
β	0	0.7	0.42	0.42
γ	0	0.0	0.40	0.40

to the classic language model (C). More importantly, it is clear that our proposed hybrid approach incorporating the semantics-based language model (C+T+S) outperforms the state-of-the-art approaches (C+T) and also (C+T+T) significantly, according to both P@10 and MAP on YA and WA.

7. CONCLUSION

In this paper, we propose a novel semantic graph based topic model (SGTM) for question retrieval. The new model supersedes the existing category-based language models because (i) question topics are more fine-grained than question categories; (ii) a question resides in only one category but could belong to multiple topics; and (iii) the integration of semantic graphs enables our topic model to capture the hidden contextual information of questions.

There are several interesting and promising directions in which this work could be extended. First, when learning the latent topics of questions, we could also include some metadata features such as the questions' categories (which have recently been shown to be useful for question retrieval [7]). Second, SGTM in the current form relies on one of the simplest topic models (PLSA), which makes sense as a first step towards integrating semantic graphs into topic models, but of course we could consider using more sophisticated topic models like LDA. Finally, besides question retrieval, SGTM could be applied to many other important tasks facing the lexical gap problem within CQA (such as question recommendation) and beyond (such as sentiment analysis on Twitter).

8. ACKNOWLEDGEMENTS

We thank the anonymous reviewer for their helpful comments. We acknowledge support from the EPSRC funded project named **A Situation Aware Information Infrastructure Project** (EP/L026015) and the **Integrated Multimedia City Data** (IMCD), a project within the ESRC-funded **Urban Big Data Centre** (ES/L011921/1). This work was also partly supported by NSF grant #61572223. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the sponsor.

Table 4: The experimental results (statistical significance using t-test: ** indicates p-value < 0.01 while * indicates p-value < 0.05).

	C	C+T	C+T+T	C+T+S
P@10 (YA)	0.265	0.324	0.336	**0.341***
MAP (YA)	0.316	0.358	0.392	**0.433****
P@10 (WA)	0.263	0.297	0.315	**0.335***
MAP (WA)	0.328	0.394	0.421	**0.453***

9. REFERENCES

[1] Yang Bao, Nigel Collier, and Anindya Datta. A partially supervised cross-collection topic model for cross-domain text classification. CIKM '13, pages 239–248.

[2] Christian Bizer, Jens Lehmann, Kobilarov, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. volume 7, pages 154–165.

[3] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. volume 3, pages 459–565.

[4] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. CIKM '08, pages 911–920.

[5] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Large-scale question classification in cqa by leveraging wikipedia semantic knowledge. CIKM '11, pages 1321–1330.

[6] Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. WWW '2010.

[7] Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. The use of categorization information in language models for question retrieval. CIKM'2009.

[8] Pak K. Chan, Martine D. F. Schlag, and Jason Y. Zien. Spectral k-way ratio-cut partitioning and clustering. DAC '93, pages 749–754.

[9] Xu Chen, Mingyuan Zhou, and Lawrence Carin. The contextual focused topic model. KDD '12, pages 96–104.

[10] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. KDD '11, pages 1271–1279.

[11] Weiwei Guo and Mona Diab. Semantic topic models: Combining word distributional statistics and dictionary definitions. EMNLP '11, pages 552–561.

[12] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. In *Machine Learning*, volume 45, pages 256–269.

[13] Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. A time-dependent topic model for multiple text streams. KDD '11, pages 832–840.

[14] Eva Hörster, Rainer Lienhart, and Malcolm Slaney. Image retrieval on large-scale image databases. CIVR '07, pages 17–24.

[15] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. Unsupervised graph-based topic labelling using dbpedia. WSDM '13, pages 465–474.

[16] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding similar questions in large question and answer archives. pages 84–90, Bremen, Germany, 2005.

[17] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. Question-answer topic model for question retrieval in community question answering. CIKM '12, pages 2471–2474, 2012.

[18] Hyungsul Kim, Yizhou Sun, Julia Hockenmaier, and Jiawei Han. Etm: Entity topic models for mining documents associated with entities. ICDM '12, pages 349–358.

[19] Fang Li, Tingting He, Xinhui Tu, and Xiaohua Hu. Incorporating word correlation into tag-topic model for semantic knowledge acquisition. CIKM '12, pages 1622–1626.

[20] Huajing Li, Zhisheng Li, Wang-Chien Lee, and Dik Lun Lee. A probabilistic topic-based ranking framework for location-sensitive domain information retrieval. SIGIR '09, pages 331–338.

[21] Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic models for word sense disambiguation and token-based idiom detection. ACL '10, pages 1138–1147.

[22] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[23] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[24] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. WWW '08, pages 101–110.

[25] Qiaozhu Mei, Deng Cai, Duo Zhang, and Chengxiang Zhai. Topic modeling with network regularization. WWW '08, pages 342–351.

[26] Michael Schuhmacher and Simone Paolo Ponzetto. Knowledge-based graph document modeling. WSDM '14, pages 543–552.

[27] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. KDD '04, pages 306–315.

[28] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. KDD '08, pages 428–437.

[29] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. WWW '08, pages 111–120.

[30] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: A rating regression approach. KDD '10, pages 783–792.

[31] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. SIGIR '06, pages 326–335.

[32] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. SIGIR '03, pages 267–273.

[33] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval models for question and answer archives. SIGIR '08.

[34] ChengXiang Zhai. *Statistical Language Models for Information Retrieval*. Morgan & Claypool Publishers, 2008.

[35] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. KDD '04, pages 743–748.

[36] Guangyou Zhou, Yang Liu, Fang Liu, Daojian Zeng, and Jun Zhao. Improving question retrieval in community question answering using world knowledge. IJCAI '13, pages 2239–2245, 2013.

[37] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. ICML '03, pages 912–919.