Latent Association Analysis of Document Pairs

Gengxin Miao[†], Ziyu Guan[†], Louise Moser[†], Xifeng Yan[†], Shu Tao[♯], Nikos Anerousis[♯], Jimeng Sun[♯]
[†]University of California, Santa Barbara
[♯]IBM T. J. Watson Research Center
[†]{miao, moser}@ece.ucsb.edu, {ziyuguan, xyan}@cs.ucsb.edu
[♯]{shutao,nikos, jimeng}@us.ibm.com

Abstract

This paper presents Latent Association Analysis (LAA), a generative model that analyzes the topics within two document sets simultaneously, as well as the correlations between the two topic structures, by considering the semantic associations among document pairs. LAA defines a correlation factor that represents the connection between two documents, and considers the topic proportion of paired documents based on this factor. Words in the documents are assumed to be randomly generated by particular topic assignments and topic-to-word probability distributions. The paper also presents a new ranking algorithm, based on LAA, that can be used to retrieve target documents that are potentially associated with a given source document. The ranking algorithm uses the latent factor in LAA to rank target documents by the strength of their semantic associations with the source document. We evaluate the LAA algorithm with real datasets, specifically, the IT-Change and the IT-Solution document sets from the IBM IT service environment and the Symptom-Treatment document sets from Google Health. Experimental results demonstrate that the LAA algorithm significantly outperforms existing algorithms.

Source Code
`http://www.uweb.ucsb.edu/~miao/resources.html`

Categories and Subject Descriptors

I.7.0 [**Computing Methodologies**]: Document and Text Processing

Keywords

Topic Model, Variational Inference, Ranking Algorithm

1. INTRODUCTION

The numerous and diverse documents generated in business and society present both challenges and opportunities for data mining research. Among the common, yet relatively unexplored, types of documents are the documents that occur in pairs. Examples of such document pairs include questions and answers, changes to IT systems and consequent problems, disease symptoms and diagnoses, *etc.* Such document pairs can be used to build valuable knowledge bases that help improve decisions in business and society.

Change (Source)	**Problem (Target)**
Set the schedule of weekly CARS backup: 3am on Sundays.	The backup is running for a long time, which is impacting the start of daytime BMP processing.
Replication of new data is loaded for all customer centers.	Server outage: User can ping the server but failed to access the database.
Back up authentication server.	User reported can access E-Pricer without inputting password.

Table 1: Example Change Problem Document Pairs.

Table 1 shows example document pairs from the IBM IT Change document sets that contain changes to IT systems (source documents) and the resulting problems (target documents). Given such document pairs, we seek to address two fundamental problems:

1. What is the underlying principle that makes the connection between a pair of documents? (*Modeling*)
2. Given a source document, how do we use this principle to rank the target documents based on how strongly they are related to the source document? (*Ranking*)

The solutions to the *Modeling* and *Ranking* problems can help us understand the semantic connection (*i.e.*, *latent association*) between paired documents and can provide tremendous value in real-world applications. For instance, in the IT service industry, changes are frequently made to an operational IT environment, and the service consultants need to evaluate the potential problems caused by a proposed change, so that they can make plans accordingly.

Both the modeling and the ranking problems present great challenges that cannot be readily addressed using existing approaches. For instance, topic models, such as CTM [4], LDA [5] and PLSI [12], are designed to model only single document sets. However, we need not only to model individual documents correctly, but also to capture the connection between the documents accurately. Furthermore, the existence of one-to-many or many-to-one mappings in a bipartite graph suggests possibly different interpretations of the topics of a document. For example, a question might refer to

KDD'12, August 12–16, 2012, Beijing, China.

different topics if the answers emphasize different aspects of the question. What we need is a model that puts a document in the context of a document pair and allows its topic proportion to be interpreted differently in different contexts. Existing topic models do not support flexible topic proportions in the same document. The ranking problem is also non-trivial. Given a source document, the number of potentially related target documents can be huge. The model needs to be able to identify an appropriate target document from a large number of candidates accurately.

In this paper, we present our novel Latent Association Analysis (LAA), which models the topic structures and their correlations together. In the LAA model, each document pair is considered as a randomly drawn correlation factor that initiates the connection between the two documents. The topic proportions of the two documents are drawn conditionally depending on the correlation factor. Each word in the documents is assumed to be generated based on a topic assignment and the topic-to-word probability distribution.

For LAA, we adopt concepts from two well-known models, namely, Correlated Topic Model (CTM) [4] and Canonical Correlation Analysis (CCA) [2]. We then develop a novel ranking algorithm to retrieve target documents based on their latent associations with the given source document. We evaluate the ranking algorithm using the IT-Change and the IT-Solution document sets from the IBM IT service environment and the Symptom-Treatment document sets from Google Health. Experimental results show that the LAA algorithm significantly outperforms existing algorithms, which confirms that LAA successfully captures the semantic-level connections among document pairs.

To the best of our knowledge, this work is the first such work to model paired document sets within a unified framework. Our contributions are two-fold. First, we show that exploring document-level correlations is more effective at capturing semantic topic associations in document pairs, compared to using word-level or topic-level correlations. LAA considers a document pair as a whole and, thus, can deliver better association semantics than existing approaches. Second, the ranking algorithm based on LAA performs exceptionally well for target document retrieval. Through diverse experiments, we show that this ranking algorithm has broad applications in document analysis and retrieval.

2. PROBLEM FORMULATION

The problem we address involves a source document set $\mathcal{D}_s$ and a target document set $\mathcal{D}_t$. Each source document $d_s \in \mathcal{D}_s$ is paired with at least one target document $d_t \in \mathcal{D}_t$, and vice versa. The pairing between source and target document sets can be represented by a bipartite graph $\mathcal{G}$, with its two sets of vertices being the source document set and the target document set, and its set of edges corresponding to the source and target document pairs. Specifically,

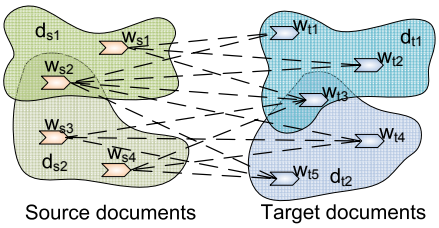
- $\mathcal{G} = \{\mathcal{D}_s \bigcup \mathcal{D}_t, \mathcal{E}\}$ is a bipartite graph with its vertices defined by a set $\mathcal{D}_s$ of source documents, a set $\mathcal{D}_t$ of target documents, and a set $\mathcal{E}$ of edges between documents in $\mathcal{D}_s$ and documents in $\mathcal{D}_t$.
- Each edge $e_i = (d_{is}, d_{it})$ represents a document pair, where $d_{is} \in \mathcal{D}_s$, $d_{it} \in \mathcal{D}_t$ and $e_i \in \mathcal{E}$.
- The vocabulary set of $\mathcal{D}_s$ is $\mathcal{W}_s = \{w_{s1}, ..., w_{sN_s}\}$, and the vocabulary set of $\mathcal{D}_t$ is $\mathcal{W}_t = \{w_{t1}, ..., w_{tN_t}\}$.

In the example in Table 1, there is a one-to-one mapping between the source and target documents. However, one-to-many or many-to-one mappings are not uncommon in other paired document sets. In this study, we consider the other mappings as special cases of one-to-one mappings and convert them to multiple one-to-one document pairs.

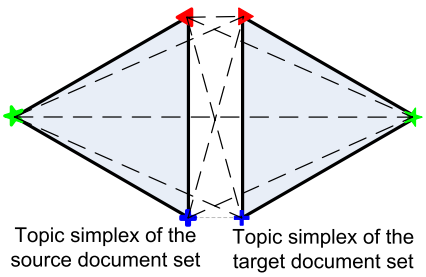
Given the above formulation, we aim to solve two problems: (1) *Modeling*: Model the associations between the source documents in $\mathcal{D}_s$ and the target documents in $\mathcal{D}_t$; (2) *Ranking*: For a new source document d_s, rank and retrieve the target document d_t, that is most likely to be associated with d_s, from a repository of target documents.

3. LATENT ASSOCIATION ANALYSIS

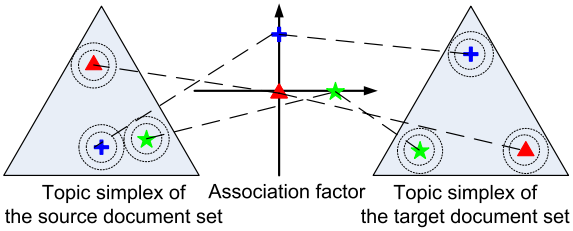
The objective of our modeling problem is different from that of existing work [4, 5, 12]. Our focus is to model the association between a pair of documents. The task is also different from traditional information retrieval tasks: (1) Our query involves a document, which is much noisier than a keyword query in traditional information retrieval tasks; (2) The source (query) document and the target documents to be retrieved arise from two separate document sets, for which we do not assume any vocabulary overlap. Therefore, similarity-based relevance scores do not apply to this problem. Conceptually, the association between the source and target documents can be considered at three different levels of granularity, yielding three possible solutions:



(a) Word-level correlation.



(b) Topic-level correlation.



(c) Document-level correlation.

Figure 1: Analyzing the Associations at Different Levels of Granularity.

Word-level correlation (Fig. 1(a)): Given individual words in the source documents, we can directly model

whether and how they are correlated with the words in the target documents using a training dataset. Unfortunately, synonyms and polysemy in free text make the correlation at the word level noisy.

Topic-level correlation (Fig. 1(b)): Topics are more stable than words. Topic-level correlation can be analyzed in two ways: 1) learn two topic structures from the two document sets separately and then discover the mapping between the two topic spaces; 2) assume that the two document sets share the same topic space and analyze the common topics. A problem with the first approach is that topics learned separately might not reflect the associations in document pairs. A problem with the second approach is that the common topic space won't be able to capture the different semantic associations between pairs of documents.

Document-level correlation (Fig. 1(c)): Instead of generating topics separately, we can learn the topics for the source and target documents simultaneously. We define a correlation factor for a document pair. The topic proportions of the two documents are drawn based on this correlation factor. In this approach, the topic distribution of each (source or target) document is studied in the context of a document pair. This approach allows flexible topic assignment if the same source document is paired with different target documents, and vice versa. Thus, the same source document can have different topic assignments in different contexts. Hence, this approach handles the one-to-many or many-to-one mappings between two document sets smoothly, because mappings involving many documents are broken down into multiple document pairs that can be considered separately.

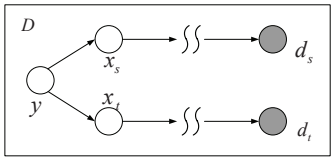


Figure 2: Basic LAA Framework.

The Latent Association Analysis (LAA) framework takes the document-level correlation approach. As shown in Fig. 2, the LAA framework consists of two components, the correlation factor y between two latent variables x_s and x_t, and the document-generation processes for d_s and d_t. We can instantiate LAA with different correlation models and topic models. Moreover, the models for generating source and target documents can differ. After learning based on training document pairs, LAA can be directly applied to solve our ranking problem: for a given query d_s, we can rank pairs (d_s, d_t) based on not only the topics of d_s and d_t, but also the correlation factor between them.

4. MODELING DOCUMENT PAIRS

In this section, we consider an instantiation of the LAA framework with the Canonical Correlation Analysis (CCA) [2] and the Correlated Topic Model (CTM) [4], and derive a variational method [3] to estimate the parameters for the model.

4.1 Canonical Correlation Analysis

Canonical Correlation Analysis [15] works on two sets of random variables and their covariance matrix. Two linear transformations are found for the two sets of random variables such that the two sets of projected variables have maximum correlation with each other. Bach *et al.* [2] gave a probabilistic interpretation of CCA and considered CCA as a model-based method that could be integrated with other probabilistic methods.

In CCA, the observed random variables $x_1 \in \mathbb{R}^{m_1}$ and $x_2 \in \mathbb{R}^{m_2}$ depend on a latent correlation factor $y \in \mathbb{R}^d$. The generative process can be described as follows.

- For a pair of variables, draw the correlation factor $y \sim \mathcal{N}(\mathbf{0}, \mathbf{I_d})$ where $\min\{m_1, m_2\} \geq d \geq 1$.
- For each set of random variables, draw
  $x_1|y \sim \mathcal{N}(T_1 y + \mu_1, \Psi_1)$, $T_1 \in \mathbb{R}^{m_1 \times d}$, $\Psi_1 \succeq 0$
  $x_2|y \sim \mathcal{N}(T_2 y + \mu_2, \Psi_2)$, $T_2 \in \mathbb{R}^{m_2 \times d}$, $\Psi_2 \succeq 0$.

In LAA, we can use CCA to capture the semantic association between the source document and the target document. The two random variables x_s and x_t are lower-dimensional representations of the source and target documents, respectively. The correlation factor y represents why these two documents are associated on a semantic level.

4.2 LAA

Whereas CCA can capture the semantic association in a document pair, other existing topic models can capture the topics of the two documents. These existing topic models include CTM [4], LDA [5], PLSI [12], *etc.* If PLSI is used, the random variables x_s and x_t are the topic proportions of the documents d_s and d_t. If LDA is used, the random variables x_s and x_t are the Dirichlet priors of the topic proportions in d_s and d_t. In both cases, the topics are assumed to be independent of each other; and the correlation between different topics cannot be properly modeled. If CTM is used, the topic proportion documents are assumed to have a multivariate Gaussian prior distribution, which is a natural fit with the Gaussian variables x_s and x_t in CCA. The correlation between different topics can be explicitly captured by the covariance matrix of the Gaussian distribution. In this paper, we choose CTM to instantiate LAA.

The instantiated LAA model is depicted in Fig. 3. The LAA model comprises the model parameters in the set $M = \{\Psi_s, T_s, \mu_s, \Psi_t, T_t, \mu_t, \beta_s, \beta_t\}$. For source and target documents d_s and d_t of lengths l_s and l_t, the words $w_{s,1:l_s}$ and $w_{t,1:l_t}$ in the source and target documents are the observable variables. The latent variables (*i.e.*, variables that are neither directly observable nor explicitly specified in the learned model) form the parameter set $V_l = \{y, x_s, x_t, z_{s,1:l_s}, z_{t,1:l_t}\}$.

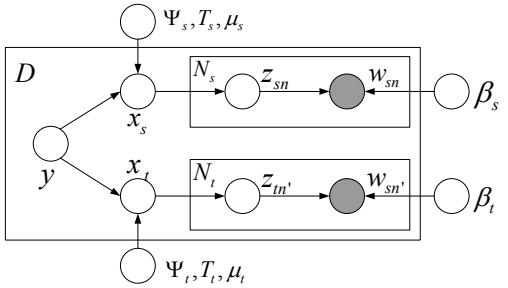


Figure 3: Instantiated LAA Model.

The generative process can be described as follows:

1. For each edge in the bipartite graph $\mathcal{G}$ (*i.e.*, a document pair), draw an L-dimensional Gaussian correlation factor: $y \in \mathcal{N}(0, I_L)$. The dimension L $<$ min $\{K_s, K_t\}$, where K_s is the number of topics in the source document set $\mathcal{D}_s$ and K_t is the number of topics in the target document set $\mathcal{D}_t$.

2. For each document pair connected by an edge, draw topic proportions as follows:
For the source document, draw
$x_s|y \sim \mathcal{N}(T_s y + \mu_s, \Psi_s)$; $T_s \in \mathbb{R}^{K_s \times L}$, $\Psi_s \succeq 0$.
For the target document, draw
$x_t|y \sim \mathcal{N}(T_t y + \mu_t, \Psi_t)$; $T_t \in \mathbb{R}^{K_t \times L}$, $\Psi_t \succeq 0$.

3. For each word in the source document, choose:
(a) a topic $z_{sn}|x_s \sim \text{Mult}(\theta_s)$, where
$\theta_{si} = \exp(x_{si})/\sum_j \exp(x_{sj})$ for $i \in \{1, 2, ...K_s\}$.
(b) a word $w_{sn}|z_{sn}, \beta_s \sim \text{Mult}(\beta_{sz_{sn}})$.
The topics and words in the target document are chosen in a similar manner.

Although the topic modeling portion of LAA stems from the idea of CTM, LAA is more complicated than the existing topic models: It is built on a set of document pairs, instead of a single document set as in existing topic models. As a result, the latent topic structures in the source and the target document sets, as well as their correlation, need to be analyzed simultaneously. LAA considers each edge in the bipartite graph as a correlation factor that initiates the connection between two documents. The generation process of the topic proportions depends on the correlation factor. That is, first, LAA decides what makes the connection between the source documents and the target documents at the document level. Then, LAA models the pair consisting of the source document and the target document as a co-occurrence interpreted by the correlation factor, instead of assuming a causal relationship between the two documents, which is difficult to validate.

It is worth noting that the topic proportion of a document is context-dependent. The same piece of text, in the eyes of different observers with different emphases, can belong to different topics. In LAA, each source or target document is put in the context of a pair, allowing the topic proportion of each document to be mutually enhanced and to be context-dependent. Doing so provides the flexibility of not deciding the topic of a document until we learn what is emphasized in the other document paired with it.

4.3 Variational Inference and Parameter Estimation

Given the LAA model described above, we solve the following problems: (1) Model fitting: Given a set of document pairs, how do we find model parameters that best fit the data? (2) Inference: For a new document pair, how do we decide the correlation factor y, the topic proportions x_s, x_t, and the topic assignments z for each word? The true posterior distributions are computationally intractable, when the hidden variables are not independent of each other, given an observed document pair. Similar to CTM, our LAA model employs a variational method to solve these problems.

4.3.1 Variational Inference

Consider a pair (d_s, d_t) of documents, represented as sets $\{w_{sn}\}$ and $\{w_{tn'}\}$ of words, where w_{sn} is the nth word in d_s and $w_{tn'}$ is the n'th word in d_t, Eq. (1) gives the probability that the document pair arises from an LAA model represented by a parameter set M.

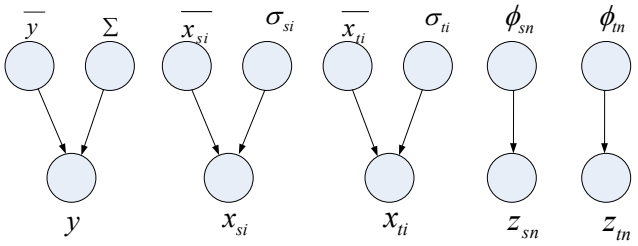


Figure 4: Variational Distribution.

$$P(d_s, d_t|M) = \int_y \int_{x_s} \int_{x_t} P(y)P(x_s|y, M)P(x_t|y, M) \times \prod_{k'=1}^{K_t} \prod_{n'=1}^{l_t} (P(z_{tn'} = k'|x_t)P(w_{tn'}|z_{tn'}, \beta_t))d(x_t) \times \prod_{k=1}^{K_s} \prod_{n=1}^{l_s} (P(z_{sn} = k|x_s)P(w_{sn}|z_{sn}, \beta_s))d(x_s)d(y) \quad (1)$$

Ideally, the latent variables in the set V_l should be chosen to maximize the probability $P(d_s, d_t|M)$ to best fit the pair of documents. Unfortunately, it is computationally intractable to determine the true posterior distribution over V_l, because the latent variables are coupled together. Thus, we introduce a variational distribution $Q(V_l)$, in which the latent variables are independent of each other, to approximate the true posterior distribution $P(V_l|d_s, d_t)$. The graphical representation of Q is shown in Fig. 4. According to the variational distribution, $Q(y) \sim \mathcal{N}(\bar{y}, \Sigma)$, $Q(x_{si}) \sim \mathcal{N}(\bar{x}_{si}, \sigma^2_{si})$, $Q(x_{ti}) \sim \mathcal{N}(\bar{x}_{ti}, \sigma^2_{ti})$, $Q(z_{sn}) \sim \text{Mult}(\phi_{sn})$ and $Q(z_{tn}) \sim \text{Mult}(\phi_{tn})$. Note that each component in the topic proportions x_s and x_t is drawn independently. The variational parameters introduced in the variational distribution are fit such that the KL-divergence between $Q(V_l)$ and $P(V_l|d_s, d_t)$ is minimized.

Using the variational distribution and Jensen's inequality, we take the logarithm of the probability in Eq. (1) and rewrite the objective function in Eq. (2). Instead of maximizing the log likelihood directly, which is intractable, we maximize the lower bound of the log likelihood to obtain an approximation of the optimal value of the latent variables.

$$\log(P(d_s, d_t|M)) \geq E_Q \log(P(d_s, d_t|M)) + H(Q) = \lfloor \mathcal{L} \rfloor \quad (2)$$

The above maximization problem is a convex optimization problem and, thus, the optimal values of the variational parameters occur when the derivatives are zero. According to the decomposition of the marginal probability in Eq. (1), we expand the lower bound of the log likelihood as follows:

$$\begin{aligned} \lfloor \mathcal{L} \rfloor = &\sum_n E_Q \log P(w_{sn}|z_{tn}, \beta_s) + \sum_{n'} E_Q \log P(w_{tn'}|z_{tn'}, \beta_t) \\ &+ \sum_n E_Q \log P(z_{sn}|x_s) + \sum_{n'} E_Q \log P(z_{tn'}|x_t) \\ &+ E_Q \log P(x_s|y, \Psi_s, T_s, \mu_s) + E_Q \log P(x_t|y, \Psi_t, T_t, \mu_t) \\ &+ E_Q \log P(y) + H(Q(V_l)) \end{aligned} \quad (3)$$

Each term on the right-hand side is a function over the variational parameters as shown in Eqs. (4) - (8):

$$\sum_n E_Q \log(P(w_{an}|z_{an}, \beta_a)) = \sum_{n=1}^{l_a}\sum_{k=1}^{K_a} \phi_{ank} \log(\beta_{ank}) \quad (4)$$

Here, a represents the source document s or the target document t in a pair. Because a document pair is symmetric, we use the same set of equations with different subscripts.

According to LAA, the topic assignment z is drawn based on the Gaussian prior x, $P(z_n = k|x) = \frac{\exp(x_k)}{\sum_j \exp(x_j)}$. Let $\iota = \sum_j \exp(x_j)$. If we take the first-order Taylor expansion with respect to ι at point ζ to approximate $\log P(z_n = k|x)$, we have $\log P(z_n = k|x) = x_k - \log(\zeta) - \frac{1}{\zeta}(\sum_j \exp(x_j) - \zeta) + \mathcal{O}((\iota - \zeta)^2)$. Thus,

$$\sum_n E_Q \log(P(z_{an}|x_a)) \geq \sum_{n=1}^{l_a}\sum_{k=1}^{K_a} \phi_{ank}\bar{x}_{ak} - l_a \log(\zeta_a) - \frac{l_a}{\zeta_a}\sum_{k=1}^{K_a} \exp(\bar{x}_{ak} + \frac{\sigma_{ak}^2}{2}) + l_a \quad (5)$$

where ζ is an additional variational parameter.

$$E_Q \log(P(x_a)) = \frac{1}{2}\log(|\Psi_a^{-1}|) - \frac{1}{2}\mathrm{tr}(\mathrm{diag}(\sigma_a^2)\Psi_a^{-1}) - \frac{1}{2}\mathrm{tr}((T_a\bar{y} + \mu_a - \bar{x}_a)(\bar{y}^T T_a^T + \mu_a^T - \bar{x}_a^T)\Psi_a^{-1}) - \frac{1}{2}\mathrm{tr}(T_a \Sigma T_a^T \Psi_a^{-1}) + \mathrm{const} \quad (6)$$

$$E_Q \log(P(y)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\mathrm{tr}(\Sigma) - \frac{1}{2}\bar{y}^T\bar{y} \quad (7)$$

$$H(Q) = -\sum_{a=s,t}\sum_{n=1}^{l_a}\sum_{k=1}^{K_a} \phi_{ank}\log(\phi_{ank}) + \frac{1}{2}\log(\det(\Sigma)) + \sum_{a=s,t}\sum_{k=1}^{K_a}\log(\sigma_{ak}) + \mathrm{const} \quad (8)$$

We substitute Eqs. (4)-(8) into Eq. (3), and then maximize the lower bound of the log likelihood by taking the partial derivatives with respect to each of the variational parameters and setting them to zero.

For the variational parameters ζ, ϕ, Σ and y, the optimal values that maximize the objective function are achieved by:

$$\zeta_a = \sum_k \exp(\bar{x}_{ak} + \frac{\sigma_{ak}^2}{2}) \quad (9)$$

$$\phi_{ank} \propto \beta_{akv}\exp(\bar{x}_{ak}) \text{ s.t. } w_{an}^v = 1 \quad (10)$$

$$\Sigma = \sum_{a=s,t} T_a^T \Psi_a^{-1} T_a + I_L \quad (11)$$

$$\bar{y} = \Sigma \sum_{a=s,t} T_a^T \Psi_a^{-1}(\bar{x}_a - \mu_a) \quad (12)$$

For the variational parameters $\bar{x}$ and σ, there are no analytical solutions. The optimal values of these variables are the solutions to Eqs. (13) and (14), which are solved iteratively using Newton's method.

$$\sum_n \phi_{an} - \frac{l_a}{\zeta_a}\exp(\bar{x}_a + \frac{\sigma_a^2}{2}) - \Psi_a^{-1}(T_a\bar{y} + \mu_a - \bar{x}_a) = 0 \quad (13)$$

$$\frac{l_a}{\zeta_a}\exp(\bar{x}_a + \frac{\sigma_a^2}{2}) + \mathrm{diag}(\Psi_a^{-1}) - \frac{1}{\sigma_a^2} = 0 \quad (14)$$

For each edge in the bipartite graph, we calculate the variational parameters using Eqs. (9) - (14) iteratively until the log likelihood lower bound in Eq. (3) no longer increases. The resulting variational parameter values are an approximation of the optimal values of the latent variables. Specifically, $y^* = \bar{y}$, $x_{ak}^* = \bar{x}_{ak}$, $z_{an}^* = \arg_k \max(\phi_{ank})$, where $a \in \{s, t\}$, $k \in \{1, 2, ..., K_a\}$, $n \in \{1, 2, ...l_a\}$.

4.3.2 Parameter Estimation

We estimate the model parameters using the variational expectation-maximization algorithm. In the E-Step, we update the variational parameters for each edge in the bipartite graph. In the M-Step, we update the model parameters, so that the sum of the log likelihood lower bound on each edge is maximized.

The process used in the M-Step is similar to that of variational inference. The goal here is to maximize the aggregated log likelihood of all the edges in the bipartite graph, rather than maximizing the log likelihood of a single edge. We sum the lower bounds of the log likelihood in Eq. (2) for each edge and take the partial derivative over the set M of model parameters. We then calculate the optimal values of the model parameters by setting the derivatives to zero.

$$\beta_{akv} \propto \sum_{e\in\mathcal{E}}\sum_n \phi_{adnk} 1(w_{ean}^v = 1) \quad \text{s.t. } \textstyle\sum_v \beta_{akv} = 1 \quad (15)$$

$$T_a = (\sum_{e\in\mathcal{E}}(\bar{x}_{ea}\bar{y}_e^T - \mu_a\bar{y}_e^T))(\sum_{e\in\mathcal{E}}(\bar{y}_e\bar{y}_e^T + \Sigma_e))^{-1} \quad (16)$$

$$\mu_a = \frac{1}{|\mathcal{E}|}(\sum_{e\in\mathcal{E}}\bar{x}_{ea} - T_a\sum_{e\in\mathcal{E}}\bar{y}_e) \quad (17)$$

$$\Psi_a = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}(\mathrm{diag}(\sigma_{ea}^2) + T_a\Sigma_e T_a^T + (T_a\bar{y}_e + \mu_a - \bar{x}_{ea})(T_a\bar{y}_e + \mu_a - \bar{x}_{ea})^T) \quad (18)$$

The E-Step and the M-Step are performed iteratively until the model parameters converge, indicating that the model parameters fit the training dataset.

5. RANKING DOCUMENT PAIRS

Given an LAA model M learned from a training dataset, for a new source document d_s, we aim to rank the target documents in a test dataset according to their potential associations with the source document. In this section, we introduce three different approaches to this problem. In Section 6, we evaluate these three approaches, together with the PTM method proposed by Zhang *et al.* [25].

5.1 Two-Step Approach

First, we discuss the Two-Step approach that mines the topics in the target and source document sets independently

and then determines the correlation between their topic structures. This method is used as the baseline for comparison with LAA.

The training process consists of two steps: (1) Find the topics in the source and target document sets; (2) Find the correlation between the source and target topic structures. In the first step, CTM is independently applied to the two document sets $\mathcal{D}_s$ and $\mathcal{D}_t$. The topic proportion priors x_s and x_t are obtained for $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively, using the variational inference method presented in [4]. For each document pair (d_s, d_t), their corresponding topic proportion priors (x_s, x_t) form a pair. In the second step, these topic proportion priors (which follow Gaussian distributions) are fed into CCA. The CCA parameters T_1, T_2, μ_1, μ_2, Ψ_1, Ψ_2 are fit to the topic proportion pairs (x_s, x_t).

In the document retrieval task, given a new source document d_s, our goal is to rank the target documents in a test set. The candidates d_t are ranked based on the probability $P(d_t|d_s)$ that a target document d_t can be observed in a document pair containing the source document d_s.

We assume that the topic proportion priors x are a lower dimensional representation of the document d. Thus, $P(d_t|d_s) \propto P(x_t|x_s)$. In CCA, given x_1, the latent correlation factor y follows a normal distribution: $y|x_1 \sim \mathcal{N}(M_1^T U_{1d}^T(x_1 - \mu_1), I - M_1 M_1^T)$ [2], and given y, x_2 follows a normal distribution: $x_2|y \sim \mathcal{N}(T_2 y + \mu_2, \Psi_2)$. Thus, given the topic proportion prior x_s of a source document d_s, its corresponding document d_t has a topic proportion prior x_t that follows a normal distribution:

$$x_t|x_s \sim \mathcal{N}(T_1 M^T(x_s - \mu_1) + \mu_2, \Psi_2 + T_1(I - MM^T)T_1^T) \quad (19)$$

where $M = (P_l)^{1/2}$ and P_l is the diagonal matrix of the top l canonical correlations.

Given a source document d_s and a candidate target document d_t, their topic proportion priors x_s and x_t can be inferred using CTM. Thus, the target documents can be ranked using $P(x_t|x_s)$, calculated from Eq. (19).

5.2 LAA Direct Approach

The LAA model presented in Section 4 allows us to predict, for a new source document d_s, which target document d_t is more likely to be associated with d_s. A direct way of ranking target documents is to evaluate how likely a hypothetical document pair (d_s, d_t) arises from the underlying LAA model. The lower bound $\log(P(d_s, d_t|M))$ can be estimated by Eq. (3) using the variational inference method discussed in Section 4.3.1. Thus, we can use the function $R(d_s, d_t) = \lfloor \log(P(d_s, d_t|M) \rfloor$ to rank the target documents. Because both the source and the target documents are considered to be a bag of interchangeable words in the LAA model, the generation probability of a long document is less than the generation probability of a short document. Note that, in this prediction method, the rank of a document pair is inversely proportional to the document length. To avoid unfairly penalizing long documents, we normalize all of the documents to unit length.

5.3 LAA Latent Approach

Although the above LAA direct approach is intuitive, it has some drawbacks. In ranking document pairs, the most important factor should be the semantic association between the source and target documents, whereas the exact wording of a document in expressing its semantic meaning should not be overemphasized. However, when evaluating a document pair using the probability that the document pair arises from the LAA model, the LAA direct approach considers all the words in the source and target documents to be equally important. As a consequence, if a target document contains rare words, it will have a low rank. The reason is that, even if the rare words in the target document might associate perfectly with the source document semantically, the probability of generating such words is still very small, which brings down the rank of the target document. Moreover, in our ranking method, the value of the correlation factor should not matter, as long as it interprets the semantic association in a document pair. The LAA direct approach cannot accommodate this feature either.

To address the aforementioned problems, we developed the LAA latent approach based on the semantic association between source and target documents. In this approach, only the topic association information is used to rank the document pairs. For any given source document d_s and candidate target document d_t, first we use variational inference to calculate the most probable correlation factor $y^* = \bar{y}$, and topic proportion $x_s^* = \bar{x}_s$ and $x_t^* = \bar{x}_t$, according to the variational distribution. Then, we evaluate how likely there exists an association between the two documents based on the topic proportion, and use the following ranking function to rank the target documents.

$$R(d_s, d_t) = P(x_s^*, x_t^*|y^*) = P(x_s^*|y^*)P(x_t^*|y^*) \quad (20)$$

In Eq. (20), $P(x_s|y^*) \sim \mathcal{N}(T_s y^* + \mu_s, \Psi_s)$, and $P(x_t|y^*) \sim \mathcal{N}(T_t y^* + \mu_t, \Psi_t)$.

6. EXPERIMENTS

We trained the LAA model based on real-world datasets and evaluated its performance for the document retrieval task. Two IT service datasets from IBM, the IT-Change and IT-Solution datasets, and a relatively smaller publicly available dataset from Google Health [1], the Symptom-Treatment dataset, were used to evaluate the effectiveness of the LAA ranking algorithm.

6.1 Datasets

The IT-Change dataset was obtained in the context of IT change management at IBM. In this dataset, each document pair consists of a change document, which describes the planned change, and a problem document, which describes the problem resulting from this change. Both the change and problem documents are written in free text, and the associations between them are established by a human expert. Given a historical IT-Change dataset, we built the LAA model and used it to retrieve the potential problem documents (from a set of problems reported) caused by a change request. The original dataset contains 24,317 pairs of documents. We randomly sampled 20,000 document pairs for training and used the rest to evaluate the performance of our ranking algorithm.

The IT-Solution dataset was obtained in the context of IT problem management at IBM. In this dataset, each document pair consists of a problem document and its corresponding solution document identified by a human expert. LAA is used to predict possible solutions for new problems. This dataset contains 19,696 pairs of documents. We ran-

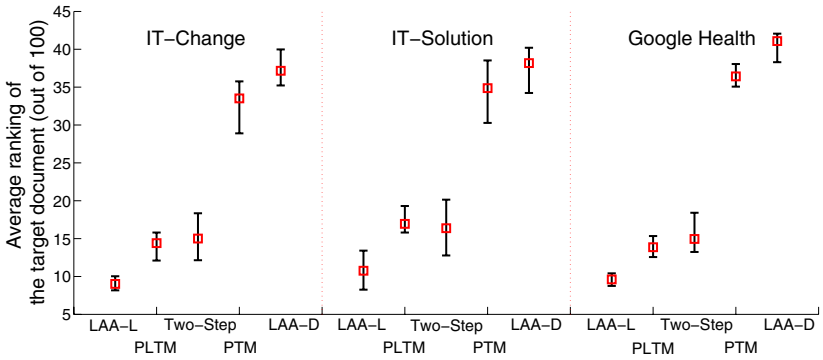


Figure 5: A Comparison of Retrieval Performance of the Different Methods Using Three Datasets.

domly selected 15,000 document pairs for training and the rest for testing.

The Google Health Symptom-Treatment dataset contains pairs of disease symptoms and treatments. This dataset is relatively small and contains 1,287 document pairs. We used 1,187 document pairs for training purposes and tested the model on the remaining 100 document pairs.

6.2 Accuracy Analysis

We compare our two LAA-based approaches, *i.e.*, LAA Direct (LAA-D) and LAA Latent (LAA-L), against the Two-Step method and the PTM method [25], as well as the Polylingual topic model (PLTM) [17], in terms of their accuracy in retrieving target documents for a given source document. For a given source document, PTM predicts a word distribution of its potential target document and compares it with the word distributions of the candidate documents. The word distribution of PTM has two components: one from the model, and the other from the similarity between the source and target documents. In LAA, we do not assume any overlap between the vocabularies of the source and target documents, which is a key advantage compared to PTM. For comparison purposes, we used only the model component in PTM as shown in Eq. (21) and adopted the KL-divergence distance [14] to evaluate the candidate target documents, as presented in [25].

$$P_{PTM}(w_t|d_s) = \sum_{i=1}^{K_s} P(w_t|\theta_i)P(\theta_i|d_s) \qquad (21)$$

Besides these methods, we also considered a two-way classification approach. We used the real source and target document pairs as positive samples, and randomly generated source and target document pairs as negative samples. Based on these labeled samples, we trained a classifier using SVM to predict future source and target document pairs. However, this approach does not work due to lack of good quality negative samples.

From each of the two datasets, we randomly selected a batch of 100 document pairs with one-to-one mappings between the source and target documents. Given a source document randomly selected within these 100 document pairs, we then ranked the 100 target documents based on the four different approaches. We used the average rank of the correct target document (the one actually paired with the selected source document) to measure the accuracy. We repeated this process for five batches (*i.e.*, 500 queries in total) for each datasets.

To train the LAA model, we varied the number of topics from 10 to 50, and chose the best performing models for the three datasets. For the IT-Change and IT-Solution datasets, we chose 20 topics for both the source and target document sets to train the three models. For the Google Health dataset, we chose 30 topics, where the model performed best. The dimensionality of the correlation factor was set to 10 for both the LAA models and the Two-Step method.

Fig. 5 compares the performance of these five approaches on the three datasets. The y-axis shows the average rank of the correct target document from the 100 target document candidates. For the IT-Change and IT-Solution datasets, each bar in the figure shows the performance range of one method over the five batches of test cases. The average over the five batches is marked in red on each bar. Because the Google Health dataset does not contain as many document pairs as the other two datasets, we investigated the accuracy of the model on only one batch of 100 test document pairs.

For all three datasets, LAA-L outperforms all other approaches, and the Two-Step method and PLTM perform closer to LAA-L. The key difference between LAA-L and the Two-Step method is that the topic structures of the source and target documents in the Two-Step method are learned independently without considering correlations between them. As a result, the performance of the Two-Step method is not as good as that of LAA-L.

PLTM, on the other hand, assumes a pair of linked documents that are identical in topic. A topic model is learned for all of the document pairs. For testing how likely a new source document and a new target document are associated, PLTM predicts their topic proportions independently and compares their similarities. PLTM is most effective when modeling the same set of articles written in different languages. In LAA, we don't assume that the topic proportions of a source document and the linked target document are the same. Hence, LAA is more general and can be applied to document sets that have arbitrary semantic associations, such as questions and answers, symptoms and diagnoses, *etc.*

LAA-D suffers from the problems discussed in Section 5.2 and does not perform well in our document retrieval task. Due to the noisy nature of word-level correlation (as noted in Section 3), the PTM method does not show good performance either. We also experimented with a modified version of PTM that compares the topic distributions, rather than the word distributions, between the source and target documents. The performance of this modified method is similar

to that of the Two-Step method, but significantly worse than that of LAA-L.

6.3 Robustness Analysis

To show the robustness of the LAA-L model in capturing the semantic associations in document pairs, we trained the model with different numbers of topics and compared the results of the document retrieval task in an experimental setting similar to that in Section 6.2.

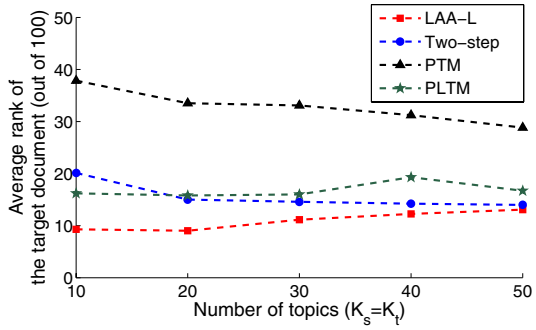


Figure 6: A Comparison of Ranks for the Different Methods Using Different Numbers of Topics.

Fig. 6 shows the experimental results for the IT-Change dataset. Limited by the space, we do not show the results for the IT-Solution and Google Health datasets, but the results are quite similar. We chose the same number of topics for the source and the target document sets, and the dimension of the correlation factor $L = \frac{1}{2}K_s = \frac{1}{2}K_t$. With different numbers of topics, the performance of the LAA-L approach remains stable, and is better than that of the other approaches.

6.4 Correlation Factor Analysis

The LAA framework assumes a correlation factor. The topic portion priors of a pair of documents are drawn centered around a point in their corresponding topic simplex. Because each point in the topic simplex implies a mixture of the topics and each topic is represented by a probability distribution of words, the point in the topic simplex can also be mapped to a distribution over words. Here, we show examples of correlation factors and the corresponding top-ranked words in the source and target documents. First, we choose a sample correlation factor, which can be mapped to the topic distributions in both the source and target document sets. The topic distribution can be further mapped to a word distribution using a linear combination of the topics. Then, we show the top-ranked word list. In these examples, the dimension of the correlation factor is set to 10. The number of topics in both the source and target document sets are set to 20. Note that the topic numbers in the source and target documents do not have to be the same.

As demonstrated in Fig. 7, the LAA model successfully captures the semantic-level connections between the source and target documents. Cases 1 and 2 were extracted from the IT-Change dataset, whereas cases 3 and 4 were from the IT-Solution dataset. For Cases 1 through 4, the top-ranked words indicate that the correlations between the source and target documents are around *Database, Network, Business,* and *Scheduling*, respectively[1].

[1]The labels for the correlation factors were added by the authors.

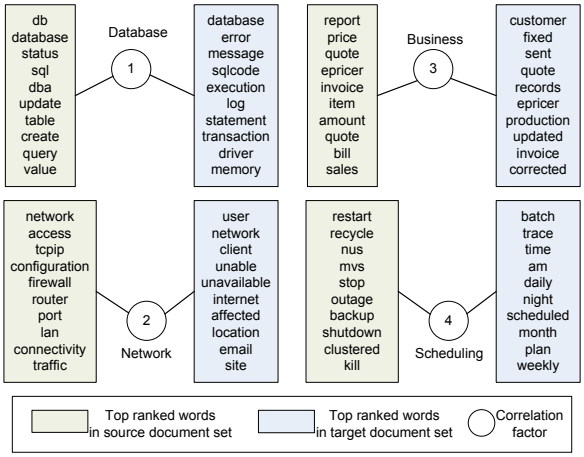


Figure 7: Sample Top Ranked Words Linked to the Same Correlation Factor.

7. RELATED WORK

Topic models have been extensively studied and have become a powerful tool for exploring the semantic content of large-scale document corpora. Most topic models deal with a single document corpus. LSI [8] uses SVD to approximate a high-dimensional document-to-word co-occurrence matrix using lower-dimensional document-to-topic and topic-to-word co-occurrence matrices. PLSI [12] introduces a probabilistic explanation of LSI. Neither LSI nor PLSI is naturally generalizable to new documents. To overcome this limitation, Blei *et al.* [5] proposed LDA, in which the topic proportions of documents are randomly drawn from a Dirichlet distribution. The Dirichlet prior is used to guide the generation of topic proportions for new documents. The CTM method [4] introduces a covariance matrix over the topic proportions and allows topics to be correlated with each other. IFTM [19] combines CTM with PCA [20] to allow exploration of a very large number of topics.

Besides the textual information in a document corpus, a number of topic models consider structural information. Steyvers *et al.* [21] use the authorship graph between authors and articles to explore author-to-topic relationships. Nallapati *et al.* [18] consider the citation graph for a document set to perform link predictions. Zhou *et al.* [28] study Web pages and tag graphs to explore user interests. Mei *et al.* [16] propose topic models with network regularization. Unlike those models, LAA focuses on document-to-document associations, and explores topics of two document sets simultaneously. Thus, LAA is better suited for ranking document pairs.

Researchers have studied topic structures of cross-lingual corpora. Zhao *et al.* [26, 27] explore probabilistic word alignments across languages using aligned bilingual document pairs, *i.e.*, the same set of articles written in two different languages. Mimno *et al.* [17] study the shared topic structure of an aligned document corpora over possibly many languages. Jagaralamudi *et al.* [13], assuming that a dictionary exists between words in two languages, analyze a single topic structure over bilingual unaligned document sets. MuTo [6] also utilizes word matchings in a dictionary to analyze the topics as distributions over the word pairs. Haghighi *et al.* [11] also applied the CCA model to learn bilingual translation lexicons.

8. CONCLUSION

This paper has presented one of the first attempts to tackle the problem of analyzing the topic structures of two document sets linked by a bipartite graph. The Latent Association Analysis (LAA) model draws the topic proportion priors of a pair of documents based on a correlation factor. Unlike other topic models, the goal of LAA is not only to provide a semantic-level explanation of the topics of the document pairs, but also to retrieve the associated target document, when a new source document is given. Using LAA, we introduced a document-level ranking method that can help to retrieve target documents associated with a source document. Experiments on real datasets confirm the effectiveness of our method for extracting semantic concepts of associated document pairs, and establish that LAA outperforms state-of-the-art algorithms in ranking document pairs. LAA can be extended to more complex association structures over multiple document sets. For other applications, the symmetric structure of the source and target documents can be replaced by an asymmetric structure, if that is more appropriate.

9. ACKNOWLEDGMENTS

The first author, Gengxin Miao, was supported by an IBM Ph.D. Fellowship. This research was also supported by the U.S. National Science Foundation under grant IIS-0954125 and by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

10. REFERENCES

[1] Google health: https://health.google.com.

[2] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Statistics Dept., UC Berkeley, 2006.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2nd ed., October 2007.

[4] D. M. Blei and J. D. Lafferty. Correlated topic models. In *NIPS*, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In *UAI*, pages 75–82, 2009.

[7] H. T. Dang, D. Kelly, and J. J. Lin. Overview of the trec 2007 question answering track. In *TREC*, 2007.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. American Society for Information Science*, 41(6):391–407, 1990.

[9] P. Forner, A. Penas, E. Agirre, I. Alegria, C. Forascu, N. Moreau, P. Osenova, P. Prokopidis, P. Rocha, B. Sacaleanu, R. Sutcliffe, and E. Tjong Kim Sang. Overview of the clef 2008 multilingual question answering track. In *Evaluating Systems for Multilingual and Multimodal Information Access*, LNCS 5706, pages 262–295. 2009.

[10] J. Gao, K. Toutanova, and W. tau Yih. Clickthrough-based latent semantic models for Web search. In *SIGIR*, pages 675–684, 2011.

[11] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779, 2008.

[12] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[13] J. Jagaralamudi and H. Daumé. Extracting multilingual topics from unaligned corpora. In *ECIR*, 2010.

[14] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR*, pages 111–119, 2001.

[15] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[16] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, pages 101–110, 2008.

[17] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. Mccallum. Polylingual topic models. In *EMNLP*, pages 880–889, Singapore, 2009.

[18] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *SIGKDD*, pages 542–550, 2008.

[19] D. P. Putthividhya, H. T. Attias, and S. Nagarajan. Independent factor topic models. In *ICML*, pages 833–840, 2009.

[20] J. Shlens. A tutorial on principal component analysis. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*, 2005.

[21] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *SIGKDD*, pages 306–315, 2004.

[22] T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *SIGIR*, pages 705–706, 2007.

[23] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS*, 2003.

[24] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *SIGIR*, pages 475–482, 2008.

[25] D. Zhang, J. Sun, C. Zhai, A. Bose, and N. Anerousis. PTM: Probabilistic topic mapping model for mining parallel document collections. In *CIKM*, pages 1653–1656, 2010.

[26] B. Zhao and E. P. Xing. Bitam: Bilingual topic admixture models for word alignment. In *COLING/ACL*, pages 969–976, 2006.

[27] B. Zhao and E. P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *NIPS*, 2007.

[28] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *WWW*, pages 715–724, 2008.