TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/10 pp 660–670
DOI: 10.26599 / TST.2018.9010011
Volume 23, Number 6, December 2018

A Latent Entity-Document Class Mixture of Experts Model for Cumulative Citation Recommendation

Lerong Ma, Lejian Liao, Dandan Song*, and Jingang Wang

Abstract: Knowledge Bases (KBs) are valuable resources of human knowledge which contribute to many applications. However, since they are manually maintained, there is a big lag between their contents and the up-to-date information of entities. Considering a target entity in KBs, this paper investigates how Cumulative Citation Recommendation (CCR) can be used to effectively detect its worthy-citation documents in large volumes of stream data. Most global relevant models only consider semantic and temporal features of entity-document instances, which does not sufficiently exploit prior knowledge underlying entity-document instances. To tackle this problem, we present a Mixture of Experts (ME) model by introducing a latent layer to capture relationships between the entity-document instances and their latent class information. An extensive set of experiments was conducted on TREC-KBA-2013 dataset. The results show that the model can significantly achieve a better performance gain compared to state-of-the-art models in CCR.

Key words: knowledge base acceleration; cumulative citation recommendation; Mixture of Experts (ME); Latent Entity-Document Classes (LEDCs)

1 Introduction

Knowledge Bases (KBs), such as Wikipedia, are widely used as reference tools to search for all kinds of information. Furthermore, they are very important in various entity-based information processing tasks, such as entity linking[1], query expansion[2, 3], knowledge graph[4], question answering[5], and entity retrieval[6]. Updating the contents of KBs is crucial to these applications. However, it is difficult for most KBs to be up-to-date because they are manually maintained by human editors. For example, there is a median time lag of one year between the publication date of a news article and the date of its publication into a Wikipedia profile[7]. This time lag would significantly decrease if documents regarding the target entity in the KBs could be detected automatically immediately the documents are published online and then recommended to the editors. This task is studied as Knowledge Base Acceleration-Cumulative Citation Recommendation (KBA-CCR) by the Text REtrieval Conference (TREC). Given a set of KB entities, KBA-CCR is used to detect relevant documents from a time-ordered corpus and evaluate their citation-worthiness to target entities.

Because of the shortage of training instances for most target entities, various global relevant supervised models (e.g., classification and learning-to-rank) have been used in the task and promising performances have been obtained[8–10]. In most models, however, all kinds of leveraged features only capture semantic and temporal information of entities and documents[11]. In fact, in our

- Lerong Ma, Lejian Liao, Dandan Song, and Jingang Wang are with Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Institute of Technology, Beijing 100081, China. E-mail: malerong@bit.edu.cn; Liaolj@bit.edu.cn; sdd@bit.edu.cn.
- Lerong Ma is also with College of Mathematics and Computer Science, Yan'an University, Yan'an 716000, China. E-mail: malerong2008@163.com.

* To whom correspondence should be addressed.

Manuscript received: 2017-03-17; revised: 2017-08-29; accepted: 2017-08-30

observation, entities and documents can provide some prior knowledge. For examples, a target entity can offer its categories or topics, and a document can offer its topics or sources. This prior knowledge embedded in the entity-document instances, referred to as class, would guide the related entity-document selection and thus impact KBA-CCR performance. For instance, when processing an entity under a category such as "politician", the KBA-CCR would probably have more preferences on a document with a political topic, but less often related to musical bands or musicians. This motivates us to leverage the prior knowledge of the entity-document instances into the Mixture of Experts (ME) model to improve KBA-CCR performance.

Mixture of experts introduced by Jacobs et al.[12] is a popular model in which different components which are "experts" can model the distribution in different regions of an input space, and the gating functions determine the probabilities of components corresponding to the regions[13]. This paper presents a Latent Entity-Document Class Mixture of Experts (LEDCME) model for KBA-CCR. Briefly, we introduce an intermediate latent layer to model Latent Entity-Document Classes (LEDCs) and define the gating functions on the observation data. We aim to achieve a mixture of experts model that can utilize prior knowledge of entity-document instances in the KBA-CCR task to improve performance.

To the best of our knowledge, this is the first research work in which prior knowledge underlying the entity-document instances is incorporated into the ME model to enhance the KBA-CCR performance. An extensive set of experiments conducted on the TREC-KBA-2013 dataset showed the effectiveness of the proposed LEDCME model.

2 Related Work

2.1 Cumulative citation recommendation

The TREC launched the KBA-CCR track from 2012 to 2014, and participants have treated CCR as either a ranking problem[14–16] or a classification problem[8, 14, 17]. Classification and learning-to-rank methods have been compared and evaluated[15, 18], and both methods can achieve competitive performances with a powerful feature set.

However, some highly supervised methods require training instances for each entity to build a relevance model, limiting their scalabilities. Entity-unspecific methods, regardless of entity distinctions, have been employed to address this problem[8, 19] with entity-document semantic and temporal features. Nevertheless, the characteristics of different entities are lost in the entity-unspecific methods. Latent entity classes have been considered[20], and they have been proven to enhance the performance. However, unlike the previous models, the proposed LEDCME enhances the ME model with latent classes information in entity-document pairs as well as entity-document semantic and temporal information.

2.2 Mixture of experts model

Mixture of experts introduced by Jacobs et al.[12] is a popular framework in the fields of machine learning to model heterogeneours data for classification, regression, and clustering[21, 22]. It has been applied to various applications in healthcare, finance, surveillance, and recognition[23].

The ME model is made up of three elements. The first one is individual component densities, which are "experts" for making predictions in their own regions. The second one is mixing coefficients known as gating functions, which determine the dominant components in a region. The last one is a probabilistic model to combine the experts and the gating functions. The experts in the ME model have been studied in classification tasks by exploiting several models, such as logistic regression[12], Support Vector Machines (SVM)[24], and multinomial[25]. In this paper, we adopt the logistic regression as experts in the paper, and similar to the conventional ME, we make use of softmax function as the gating function in our LEDCME model.

3 Mixture of Experts Model for CCR

In this section, a novel learning framework is proposed for CCR, which is an ME model that combines logistic regression as experts and softmax function as gating functions. The gating function models the LEDCs, and the logistic regression models the relevance of the entity-document instances. We first define the research problem and model it as a classification task, and then propose LEDCME model for CCR. Finally, we estimate the model parameters by using the log-likelihood loss function and Expectation-Maximization (EM) algorithm.

3.1 Problem statement

We consider CCR as a binary classification problem that treats the relevant entity-document pairs as positive instances and irrelevant ones as negative instances.

Given a set of KB entities $\mathcal{E} = \{e_u|u = 1,\ldots,M\}$

and a document collection $\mathcal{D} = \{d_v | v = 1, \ldots, N\}$, our objective is to estimate the relevance of a document d to a given entity e. In other words, we need to estimate the conditional probability of relevance $P(r|e,d)$ with respect to an entity-document instance (e,d), where $r \in \{-1,+1\}$. When $r = +1$, it indicates a positive instance, otherwise, a negative instance. Given an entity-document instance, we consider two kinds of features. One is for the features extracted from the entity and the document that are represented as a feature vector $f(e,d) = (f_1(e,d), \ldots, f_K(e,d))$, where K indicates the number of entity-document features. The other is for the LEDC information that are represented as a feature vector $g(e,d) = (g_1(e,d), \ldots, g_L(e,d))$, where L denotes the number of entity-document classes features. The entity-document features and the entity-document class features are introduced in Section 4.

3.2 Entity-document class mixture of experts model

The ME model has been applied to classification tasks[23]. "Experts" can model a distribution in different regions of input space, and the gating functions weight the relevance of the experts. As we present the LEDC information, different LEDCs should correspond to different classifiers to improve the classification performance. Presumably, the ME model is suitable for the above cases; therefore, we apply it to the CCR task with the following problem formulation.

Given (e,d) denoting an entity-document instance with a target relevant level $r \in \{-1,1\}$, we introduce a variable $z \in \{1,2,\ldots,N_z\}$ as experts to capture the LEDC information where N_z is the number of experts, and define

$$P(z=k|(e,d);\alpha) = \frac{\exp(b_k + \sum_{j=1}^{L} \alpha_{kj} g_j(e,d))}{\sum_{h=1}^{N_z} \exp(b_h + \sum_{j=1}^{L} \alpha_{hj} g_j(e,d))} \tag{1}$$

where $g_j(e,d)$ is the weight for the j-th entry of the entity-document class information vector $g(e,d)$, b_k is a bias parameter of the k-th entity-document class, α_k is the L-dimensional coefficients vector associated with z, α_{kj} is the j-th entry of the vector of α_k, and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{N_z})$ is the parameter vector for the multinomial logistic model with softmax functions. Equation (1) corresponds to the gating function representing the probability of the k-th LEDC. For simplicity, we define an additional dummy feature $g_0(e,d) = 1$ and let $\alpha_{k0} = b_k$; then Eq. (1) can be written in a form that

$$P(z=k|e,d;\alpha) = \frac{1}{Z} \exp(\sum_{j=0}^{L} \alpha_{kj} g_j(e,d)) \tag{2}$$

where $Z = \sum_{h=1}^{N_z} \exp(\sum_{j=0}^{L} \alpha_{hj} g_j(e,d))$. Next, we define

$$P(r=1|e,d,z;\omega) = \delta\left(\sum_{i=0}^{K} \omega_{zi} f_i(e,d)\right) \tag{3}$$

Equation (3) denotes the z-th expert corresponding to a logistic regression model under the z-th LEDC, where ω_{zi} is the weight for the i-th feature vector entry for the given training instance (e,d) under the hidden class z, $f_0(e,d) = 1$ is a dummy feature, $\omega_z = (\omega_{z1}, \ldots, \omega_{zK})$, $\omega = (\omega_1, \omega_2, \ldots, \omega_{N_z})$ is a vector of parameters for all experts, and $\delta(\cdot)$ is the sigmoid function. From Eq. (3), we can derive that

$$P(r=-1|e,d,z;\omega) = 1 - \delta\left(\sum_{i=0}^{K} \omega_{zi} f_i(e,d)\right) = \delta\left(-\sum_{i=0}^{K} \omega_{zi} f_i(e,d)\right) \tag{4}$$

According to Eqs. (3) and (4), the general representation of the experts is given by

$$P(r|e,d,z;\omega) = \delta\left(r \sum_{i=0}^{K} \omega_{zi} f_i(e,d)\right) \tag{5}$$

Finally, we combine the gating function (Eq. (2)) and the expert function (Eq. (5)), and obtain the LEDCME written in the form as follows:

$$P(r|e,d;\alpha,\omega) = \frac{1}{Z} \sum_{z=1}^{N_z} \exp(\sum_{j=0}^{L} \alpha_{zj} g_j(e,d)) \delta\left(r \sum_{i=0}^{K} \omega_{zi} f_i(e,d)\right) \tag{6}$$

where N_z is the number of experts corresponding to the number of the LEDCs.

3.3 Model parameter estimation

We use maximum likelihood to determine the parameters (i.e., α and ω) of the ME model.

Suppose we have a dataset of entity-document observations represented as $\mathcal{T} = \{(e_u, d_v) | u = 1, \ldots, M; v = 1, \ldots, N\}$ and $\mathcal{R} = \{r_{uv} | u = 1, \ldots, M; v = 1, \ldots, N\}$ denotes the corresponding relevance judgement (i.e., $+1$ or -1), and we aim to generate this data using LEDCME (Eq. (6)). Assuming that entity-document observations $\mathcal{T}$ are drawn independently from the distribution, according to Eq. (6), the likelihood function is given by

$$P(\mathcal{R}|\alpha,\omega) = \prod_{u=1}^{M} \prod_{v=1}^{N} P(r_{uv}|e_u,d_v) = \prod_{u=1}^{M} \prod_{v=1}^{N} \left(\frac{1}{Z} \sum_{z=1}^{N_z} \exp(\sum_{j=0}^{L_z} \alpha_{zj} g_j(e_u,d_v)) \delta(r_{uv} \sum_{i=0}^{K} \omega_{zi} f_i(e_u,d_v)) \right) \tag{7}$$

Traditionally, we define the log-likelihood loss function in the form

$$E(\alpha,\omega) = -\mathtt{ln}P(\mathcal{R}|\alpha,\omega) \tag{8}$$

Note that the log-likelihood loss function can exhibit severe over-fitting for the data set $\mathcal{T} = \{(e_u,d_v)|u = 1,\ldots,M; v = 1,\ldots,N\}$ when the dataset is linearly separable. In such cases, the over-fitting phenomenon is often controlled by adding a regularization term to the error function. Here, we adopt the $L2$ regularization method which takes the form of a sum of squares of all of the coefficients. This leads to a modified error function of the form.

$$E(\alpha,\omega) = -\mathtt{ln}P(\mathcal{R}|\alpha,\omega) + \lambda\|(\alpha,\omega)\|_2^2 \tag{9}$$

Here, the coefficient λ governs the relative importance of the regularization term and the log-likelihood loss function term, and (α,ω) is the vector of all parameters of the model defined in Eq. (6) that will be learned.

The object function Eq. (9) contains latent variables (i.e., the hidden entity-document class z), a typical approach to minimize the object function is to use the EM algorithm[26] by E-step and M-step iterations until convergence. Here we have to point out that the standard EM algorithm is to maximum the log-likelihood function, while the loss function Eq. (9) is used to minimize the negative log-likelihood function; therefore, both methods are equivalent. In addition, the optimization Eq. (9) of E-step is the same as the standard EM algorithm, because the distribution $Q(z)$ defined over the latent variables does not appear in the regularization term. Moreover, the M-step typically requires only a small modification to the M-step of the standard EM algorithm. The detailed derivation of the variant EM can be seen in Ref. [13].

The E-step can be derived by computing the posterior probability of z given α and ω for an entity-document pair (e_u,d_v).

$$P(z|e_u,d_v) = \frac{\exp(\sum_{j=0}^{L_z}\alpha_{zj}g_j(e_u,d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zi}f_i(e_u,d_v))}{\sum_z\exp(\sum_{j=0}^{L_z}\alpha_{zj}g_j(e_u,d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zi}f_i(e_u,d_v))} \tag{10}$$

According to the EM algorithm, the variant Q function of Eq. (9) is the following:

$$Q([\alpha,\omega],[\alpha,\omega]^{\text{old}}) = -\sum_{uv}\sum_{z}P(z|e_u,d_v)\cdot \left[\log\left(\delta(r_{uv}\sum_{i=0}^{K}\omega_{zi}f_i(e_u,d_v))\right) + \log\left(\frac{1}{Z}\exp(\sum_{j=0}^{L_z}\alpha_{zj}g_j(e_u,d_v))\right)\right] + \lambda\|(\alpha,\omega)\|_2^2 \tag{11}$$

Therefore, we can get the following parameters update rules for the M-step.

$$\omega_z^* = \arg\min_{\omega_z} -\sum_{uv}P(z|e_u,d_v)\cdot \log\left(\delta(r_{uv}\sum_{i=0}^{K}\omega_{zi}f_i(e_u,d_v))\right) + \lambda\|\omega_z\|_2^2 \tag{12}$$

and

$$\alpha_z^* = \arg\min_{\alpha_z} -\sum_{uv}P(z|e_u,d_v)\cdot \log\left(\frac{1}{Z}\exp(\sum_{j=0}^{L_z}\alpha_{zj}g_j(e_u,d_v))\right) + \lambda\|\alpha_z\|_2^2 \tag{13}$$

To optimize Eqs. (12) and (13), we utilize the `minFunc` toolkit[27] by employing Quasi-Newton strategy. The hyper-parameters N_z and λ are determined by using cross-validation.

The LEDCME has two advantages against the logistic regression. One is that the combination parameters vary across various entity-document classes and this variation leads to a gain of flexibility, and the other is that it offers probabilistic semantics for latent entity-document classes and thus entity-document pairs can be associated with multiple classes.

3.4 Two special cases of LEDCME

Based on the previous proposed LEDCME model, we can realize a Latent Entity Class Mixture of Experts (LECME) model when the LEDCME model is simplified with a single entity class. Thus the LECME model is a special case of the LEDCME model. The major difference is that LECME utilizes the entity class feature vector for the gating function rather than the entity-document class feature vector. Similar to the LEDCME model, the LECME model is illustrated as follows.

Given (e,d) denoting an entity-document instance and a target relevant level $r \in \{-1,1\}$, we introduce a variable $z \in \{1,2,\ldots,N_z\}$ to capture the latent entity class information where N_z is the number of experts, and define a probability distribution as the gating function in the following:

$$P(z=k|e,d;\alpha) = \frac{1}{Z_e}\exp(\sum_{j=0}^{A}\alpha_{kj}g_j(e)) \tag{14}$$

where $Z_e = \sum_{h=1}^{N_z}\exp(\sum_{j=0}^{A}\alpha_{hj}g_j(e))$, $g_0(e) = 1$ is a dummy element of the entity class feature vector, and $\alpha = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kA})$ is a vector of parameters for the gating function. We take Eq. (5) as the expert funltion in the following.

$$P(r|e,d,z;\omega) = \delta\left(r\sum_{i=0}^{K}\omega_{zi}f_i(e,d)\right) \tag{15}$$

Finally, we combine the gating function (Eq. (14)) and the expert (Eq. (15)), and obtain the LECME model written in the following form:

$$P(r|e,d;\alpha,\omega) = \\ \frac{1}{Z_e}\sum_{z=1}^{N_z}\exp(\sum_{j=0}^{A}\alpha_{zj}g_j(e))\delta\left(r\sum_{i=0}^{K}\omega_{zi}f_i(e,d)\right) \tag{16}$$

where N_z is the number of experts corresponding to the number of the latent entity classes.

Similarly, we can obtain another special case of LEDCME with a single document class, called as Latent Document Class Mixture of Experts (LDCME) model, and the representation of LDCME is as follows:

$$P(r|e,d;\alpha,\omega) = \\ \frac{1}{Z_d}\sum_{z=1}^{N_z}\exp(\sum_{j=0}^{B}\alpha_{zj}g_j(d))\delta\left(r\sum_{i=0}^{K}\omega_{zi}f_i(e,d)\right) \tag{17}$$

Here, N_z is the number of experts corresponding to the number of the latent document classes.

4 Features

In this section, the two kinds features used in our LEDCME model are presented. Entity-document features (i.e., $f(e,d)$) were employed in the experts presented in Eq. (5). Moreover, LEDCME needs entity-document class features (i.e., $g(e,d)$) to learn the gating functions that correspond to the LEDCs information.

Since our aim is not to develop new entity-document features, we employed the same entity-document feature set presented in previous studies[8, 19], which were used effectively. They are listed in Table 1.

Table 1 The features of entity-document pairs.

Feature	Description
$N(e_{\text{rel}})$	# Entity e's related entities found in its profile page
$N(d,e)$	# Occurrences of e in document d
$N(d,e_{\text{rel}})$	# Occurrences of the related entities in document d
$\text{FPOS}(d,e)$	First occurrence position of e in d
$\text{FPOS}_n(d,e)$	$\text{FPOS}(d,e)$ normalized by the document length
$\text{LPOS}(d,e)$	Last occurrence position of e in d
$\text{LPOS}_n(d,e)$	$\text{LPOS}(d,e)$ normalized by the document length
$\text{Spread}(d,e)$	$\text{LPOS}(d,e) - \text{FPOS}(d,e)$
$\text{Spread}_n(d,e)$	$\text{Spread}(d,e)$ normalized by document length
$\text{Source}(d)$	The source of d
$\text{weekday}(d)$	Weekday of d published
$\text{burst}(d)$	Burst weights of d

According to the entity-document class information, we considered two groups of prior knowledge: the prior knowledge of entities and the prior knowledge of documents. Finally, we combined these two groups of prior knowledges to produce the entity-document class information.

4.1 Prior knowledge of entities

We considered two types of prior knowledge of entities to develop entity-related features.

Profiled features. Every entity in KBs, such as Wikipedia and Twitter, has a unique profile page that contains its basic information of this entity, including name, location, and biography. The profile pages of all target entities were acquired from the Wikipedia and Twitter as a profile collection. After preprocessing of the profile collection by removing stop words and stemming, we used the bag-of-words model to represent each target entity as a vector, where term weights were determined by Term Frequency-Inverse Document Frequency (TFIDF) scheme.

Category features. Some KBs such as Wikipedia curate entities using hierarchical categories. For instance, Blair Thoreson in Wikipedia is labelled with categories including a member of the North Dakota House of Representatives, 1964 births, living people, politicians from Fargo, and North Dakota. We append three meta-categories: person, organization, and facility that cover all the entities in the target entity set. Like profile features, we leveraged bag-of-categories to represent the categories of the entity as a vector of features, where category weights were designated 1 if the specific category is occurrent, and 0 otherwise.

4.2 Prior knowledge of documents

Topic-based features. The prior knowledge underlying a document is its intrinsic topics. We modeled topics of documents by adopting bag-of-words and Latent Dirichlet Allocation (LDA) models. After removing stop words and stemming, we useed the gensim[28] package to generate the vector of documents by employing the bag-of-words model, where term weights were determined by TFIDF scheme. In addition, we used JGibbLDA[29], which is a java implementation of LDA that uses Gibbs sampling for parameter estimation and inference, to produce the vector of topics of a document in the dataset. Consequently, two kinds of features for the topic of a document were produced: TFIDF-based features and LDA-based features.

Source-based features. Another prior knowledge of a document is its source to evaluate the probability of the document's reliability. For example, a document from the news of the government is more reliable than a document from web chat. We leveraged a "bag-of-sources" model to represent each document as a feature vector, and term weights were determined in terms of a binary occurrence scheme.

5 Experimental Section

5.1 Dataset

We conducted experiments on the TREC-KBA-2013 dataset[30], which consists of a temporally stream corpus and a target entity set. The stream corpus comprised roughly 1 billion documents culled from 10 sources including news, social, weblog, and so on. The stream corpus is divided into the training data with documents from October 2011 to February 2012 and the testing data with other documents. We followed this convention in our experiments. The target entity set was composed of 121 Wikipedia entities and 20 Twitter entities.

Each entity-document instance was assessed as one of four-point rating levels. (1) Vital. This denotes timely information of the entity's current state, actions, or situation of the entity. This motivates a change to the entity's profile. (2) Useful. This denotes background information, such as biography and secondary source information. (3) Neutral. This denotes informative but not citation-worthy information, e.g., tertiary source such as Wikipedia articles. (4) Garbage. This denotes information from which noting about the target entity can be learned from the document, e.g., spam. The detailed annotation of the dataset is listed in Table 2.

5.2 Evaluation scenarios

According to different granularity settings and the target of the CCR task, we evaluate the proposed models in two classification scenarios respectively.

Vital Only. Only vital entity-document pairs are treated as positive instances, and the others as negative instances. This scenario is the essential task of CCR.

Table 2 Detailed annotation of the dataset.

	Rating level				
	Vital	Useful	Neutral	Garbage	Total
Training set	1 696	2 121	1 030	1 702	6 549
Test set	5 630	11 579	3 379	10 543	31 131

Vital+Useful. Both vital and useful entity-document pairs are treated as positive instances, and the others as negative ones.

5.3 Experimental setting

We carried out the experiments on a 64-bit machine with Intel Xeon 2.4 GHz (L5530), 4 MB cache, and 24 GB memory. The loss objective function Eq. (9) involves two hyper-parameters: One is the number of LEDCs N_z with regard to the number of experts, and the other is λ governed tradeoff between the error loss function and the regularization term. In this study, a 5-fold cross-validation was utilized to select the two hyper-parameters on a grid (N_z,λ), where $N_z \in \{2,3,\ldots,50\}$ and $\lambda \in \{\exp(-50),\exp(-49),\ldots,\exp(0)\}$.

5.4 Experimental methodology

Experiments using nine variants of LEDCME were conducted on the TREC-KBA-2013 dataset. For further comparison with LEDCME, we also conducted experiments related to LECME and LDCME by replacing the entity-document class information with only entity class information and document class information, respectively by setting the other one as only one class.

5.4.1 Latent entity class mixture of experts models

- Profile-based entity class ME model (Profile_LECME). This is a variant of LECME that utilizes profile-based features as entity class features for the gating function.

- Category-based entity class ME model (Category_LECME). This is a variant of LECME that utilizes category-based features as entity-class features for the gating function.

- Combine entity class ME model (Combine _LECME). This is a variant of LECME that utilizes profile-based and category-based entity features together as entity class features for the gating function. In our experimental setting, we simply combine the two types of entity class feature vectors together into an integral feature vector.

5.4.2 Latent document class mixture of experts models

- Source-based document class ME model (Source_LDCME). This is a variant of LDCME that uses source-based features as document class features for the gating function.

- TFIDF-based document class ME model (TFIDF_

LDCME). This is a variant of LDCME that uses TFIDF-based features as document class features for the gating function.

- LDA-based document class ME model (LDA_LDCME). This is a variant of LDCME that employs LDA-based features as document class features for the gating function.

5.4.3 Entity-document class mixture of experts

(1) Profile catenated with document information.

- Profile+Source-based LEDCME (Profile+Source_LEDCME). This is a variant of LEDCME that utilizes profile features of entities catenating source features of documents as the entity-document class features for the gating function.

- Profile+TFIDF-based LEDCME (Profile+TFIDF_LEDCME). This is a variant of LEDCME that uses profile features of entities catenated with TFIDF features of documents as the entity-document class features for the gating function.

- Profile+LDA-based LEDCME (Profile+LDA_LEDCME). This is a variant of LEDCME that uses profile features of entities catenated with LDA features of documents as entity-document class features for the gating function.

(2) Category catenated with document information.

- Category+Source-based LEDCME (Category+Source_LEDCME). This is a variant of LEDCME that utilizes category features of entities catenated with source features of documents as the entity-document class features for the gating function.

- Category+TFIDF-based LEDCME (Category+TFIDF_LEDCME). This is a variant of LEDCME that uses category features of entities catenated with TFIDF features of documents as the entity-document class features for the gating function.

- Category+LDA-based LEDCME (Category+LDA_LEDCME). This is a variant of LEDCME that uses category features of entities catenated with LDA features of documents as entity-document class features for the gating function.

(3) ProCat catenated with document information.

- ProCat+Source-based LEDCME (ProCat+Source_LEDCME). This is a variant of LEDCME that utilizes ProCat features of entities catenated with source features of documents as the entity-document class features for the gating function, where we appended the profile and category features together into an integral features as ProCat features of entities.

- ProCat+TFIDF-based LEDCME (ProCat+TFIDF_LEDCME). This is a variant of LEDCME that uses ProCat features of entities catenated with TF-IDF features of documents as the entity-document class features for the gating function. We appended the profile and category features together into an integral features as ProCat features of entities.

- ProCat+LDA-based LEDCME (ProCat+LDA_LEDCME). This is a variant of LEDCME that uses ProCat features of entities catenated with LDA features of documents as entity-document class features for the gating function. We appended the profile and category features together into integral features as ProCat features of entities.

For reference, we also include three top-ranked approaches in the TREC-KBA-2013 track, and the logistic regression model as our baselines.

- Official baseline[10]. This is an official baseline in which the annotators manually select a list of keywords of the target entities for filtering vital and useful documents.

- BIT-MSRA[8]. This is an entity-unspecific random forests classification model with the first place approach in TREC-KBA-2013 track. In this approach, 13 types of features are extracted between entities and documents, and a global model is learned for all entities using the random forest classification model.

- UDEL[9]. This is an entity-centric query expansion approach that achieves the second performance in TREC-KBA-2013 track. In this approach, related entities are first detected from the profile page of a given target entity. Then, the target entity combines the related entities as new queries and ranks the relevant detected documents.

- Logistic Regression (LR). This is the LR model on the TREC-KBA-2013 dataset.

5.5 Overall results

We adopted precision P, recall R, and harmonic mean $F1$ (harmonic mean between precision and recall) as the evaluation measurements. All the measurements were computed in an entity-insensitive manner, that is, they were computed based on the test pool of all entity-document pairs regardless of specific entities. Furthermore, low recall and high precision would lead to the manual inspection of fewer documents, and important documents may be missed. On the other hand, high recall and low precision would lead to the review of more documents, which may not be feasible if editors are limited. Therefore, we focus on $F1$ measurements in this study.

The overall results on the TREC-KBA-2013 dataset are given in Table 3. Compared with all the baselines listed in the 2nd block of Table 3, our LEDCME models and the simplified LECME and LDCME models achieved significantly higher or competitive $F1$ in both scenarios.

All variants of LECME outperformed all the baselines in both the scenarios. In particular, Combine_LECME achieved significantly better $F1$ performance in both the scenarios. This means that the profile and category information can enhance each other when used as the entity class information.

All variants of LDCME yielded better $F1$ performance in both the scenarios. However, Source_LDCME performed poorly in the two scenarios. This intuitively demonstrates that document source is not a crucial factor in determining the importance of documents.

Among the LEDCME models, the ProCat+TFIDF_ LEDCME model achieved the best $F1$ value in the Vital Only scenario, which improved $F1$ by about 47% relative to the official baseline. Its $F1$ value also exceeded those of the other comparative models. In the Vital+Useful scenario, the Profile+TFIDF_LEDCME model achieved the best $F1$ value among the LEDCME models, which increased by approximately 7% relative to the official baseline. In both scenarios, all the variants of LEDCME also outperformed the LR model. In comparison to LR, our best model improved $F1$ by about 41% and 7% in the Vital and Vital+Useful scenarios, respectively. These comparisons clearly show the overall effectiveness of our LEDCME model.

Moreover, our LEDCME models outperformed LECME and LDCME approaches in Vital Only scenarios. In comparison with seperately Combine_LECME (best among LECME models) and TFIDF_LDCME, their combination, ProCat+TFIDF_ LEDCME model, achieved the highest $F1$ values by improving the $F1$ values of Combine_LECME and TFIDF_LDCME by 5% and 26%, respectively. This indicates that the latent class information in entity-document pairs is more useful than the separate latent class information in entities and documents in the Vital Only scenario. Similar phenomena was exhibited in other combinations. For example, compared with Category_LECME and TFIDF_LDCME, Category+ TFIDF_LEDCME increased $F1$ by 18% and 25%, respectively. Moreover, in contrast with Profile_LECME and TFIDF_LDCME, Profile+TFIDF_LEDCME increased $F1$ by 21% and 25%, respectively. These results validate our motivations that (1) incorporating the LEDCs information in ME can enhance citation recommendation quality, and (2) profile and category features of entities and TFIDF or LDA features of documents can capture the LEDCs information.

Furthermore, all variants of LEDCME with regard to document source performed worse than those in the Vital

Table 3 Overall results of evaluated models on the TREC-KBA-2013 dataset.

Method	Vital Only			Vital + Useful		
	P	R	$F1$	P	R	$F1$
Official Baseline	0.171	**0.942**	0.290	0.540	**0.972**	0.694
BIT-MSRA	0.214	0.790	0.337	0.589	0.974	0.734
UDEL	0.169	0.806	0.280	0.573	0.893	0.698
LR	0.218	0.507	0.304	0.604	0.913	0.727
Profile_LECME	0.332	0.376	0.353	0.669	0.866	0.755
Category_LECME	0.316	0.422	0.362	0.672	0.894	0.767
Combine_LECME	0.397	0.418	0.407	0.703	0.877	**0.780**
Source_LDCME	0.286	0.230	0.255	0.615	0.851	0.714
TFIDF_LDCME	0.313	0.379	0.343	0.712	0.839	0.769
LDA_LDCME	0.396	0.341	0.366	**0.734**	0.828	0.778
Profile+Source_LEDCME	0.250	0.621	0.356	0.640	0.886	0.743
Profile+TFIDF_LEDCME	**0.405**	0.449	0.426	0.681	0.898	0.774
Profile+LDA_LEDCME	0.331	0.584	0.422	0.639	0.870	0.737
Category+Source_LEDCME	0.281	0.478	0.354	0.628	0.909	0.744
Category+TFIDF_LEDCME	0.403	0.454	0.427	0.674	0.903	0.771
Category+LDA_LEDCME	0.361	0.497	0.418	0.631	0.922	0.749
ProCat+Source_LECDME	0.311	0.429	0.361	0.631	0.909	0.745
ProCat+TFIDF_LEDCME	0.398	0.462	**0.428**	0.685	0.882	0.772
ProCat+LDA_LEDCME	0.404	0.416	0.410	0.646	0.892	0.749

Only scenario. These results agree with our previous discussion about the futility of document sources. Topic-based features of documents, including TFIDF and LDA, have far more dimensions than source-based features of documents. However, although source-based features of documents have only small dimensions (10 in our experiments), Profile+Source_LEDCME, Category+Source_LEDCME, and ProCat+Source_LECDME achieved better $F1$ than LR in the Vital Only scenario. Therefore, the performance can be boosted further if we can design more valuable features to represent the entity-document classes information.

Moreover, the $F1$ differences among Profile+TFIDF_LEDCME, Category+TFIDF_LEDCME, and ProCat+TFIDF_LEDCME are marginal in both the scenarios, and the $F1$ differences among Profile+LDA_LEDCME, Category+LDA_LEDCME, and ProCat+LDA_LEDCME are also small in both the scenarios. These results show that the strategies in which entities are catenated with document class information. This motivates us to develop further better combination strategies to improve the performance of CCR.

In the Vital+Useful scenario, the Combine_LECME model achieved the highest $F1$ value. However, there is little difference between the $F1$ values of the LECME, LDCME, and LEDCME models. This is probably because the Vital+Useful scenario is not an important task that there are some disagreements in the annotation data.

6 Conclusion

The objective of CCR is to filter citation-worthy documents for a set of KB entities from a chronological stream corpus. To address the problem of training data insufficiency for entities, we propose the LEDCME by utilizing latent class information in entity-document pairs, with the profiles and categories of the target entities, as well as topics features of documents, TFIDF and LDA. We conducted extensive experiments on the TREC-KBA-2013 dataset, and the results demonstrate that (1) when introducing the latent entity-document information, the ME models are effective for CCR, (2) profiles and categories of entities and topics and TFIDF of documents can capture the entity-document class information, and (3) strategies in which entity is catenated with document information are effective combination strategies.

For our future work, we plan to explore more useful entity-document class information, and apply it to more proper combination strategies between latent entity classes and document classes to improve CCR performance.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2016YFB1000902), the National Natural Science Foundation of China (Nos. 61472040, 61751217, and 61866038), Natural Science Basic Research Plan in Shaanxi Province of China (No. 2016JM6082), and PhD start project of Yan'an University (No. YDBK2018-09).

References

[1] R. Mihalcea and A. Csomai, Wikify!: Linking documents to encyclopedic knowledge, in *Proc. 16th ACM Conf. Conf. Information and Knowledge Management*, Lisbon, Portugal, 2007, pp. 233–242.

[2] Y. Xu, G. J. F. Jones, and B. Wang, Query dependent pseudo-relevance feedback based on Wikipedia, in *Proc. 32nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Boston, MA, USA, 2009, pp. 59–66.

[3] J. Dalton, L. Dietz, and J. Allan, Entity query feature expansion using knowledge base links, in *Proc. 37th Int. ACM SIGIR Conf. Research & Development in Information Retrieval*, Queensland, Australia, 2014, pp. 365–374.

[4] C. H. Zhang, M. Zhou, X. Han, Z. Hu, and Y. Ji, Knowledge graph embedding for hyper-relational data, *Tsinghua Sci. Technol.*, vol. 22, no. 2, pp. 185–197, 2017.

[5] H. T. Dang, D. Kelly, and J. Lin, Overview of the TREC 2007 question answering track, in *Proc. 16th Text Retrieval Conf.*, Gaithersburg, MD, USA, 2007, pp. 1–18.

[6] K. Balog, P. Serdyukov, and A. P. de Vries, Overview of the TREC 2010 entity track, in *Proc. 19th Text Retrieval Conf.*, Gaithersburg, MD, USA, 2010.

[7] J. R. Frank, M. K. Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Ré, and I. Soboroff, *Building An Entity-centric Stream Filtering Test Collection for TREC 2012*. Massachusetts Institute of Technology, 2012.

[8] J. G. Wang, D. D. Song, C. Y. Lin, and L. J. Liao, Bit and MSRA at TREC KBA CCR track 2013, in *Proc. 22nd Text Retrieval Conf.*, Gaithersburg, MD, USA, 2013.

[9] X. T. Liu, J. Darko, and H. Fang, A related entity based approach for knowledge base acceleration, in *Proc. 22nd Text Retrieval Conf.*, Gaithersburg, MD, USA, 2013.

[10] J. R. Frank, S. J. Bauer, M. K. Weiner, D. A. Roberts, N. Tripuraneni, C. Zhang, and C. Ré, Evaluating stream filtering for entity profile updates for TREC 2013, in *Proc. 22nd Text Retrieval Conf.*, Gaithersburg, MD, USA, 2013.

[11] L. R. Ma, D. D. Song, L.J. Liao, and J. G. Wang, PSVM: A preference enhanced SVM model using preference data for classification, *Science China Information Sciences*, vol. 60, no. 12, pp. 1–14, 2017.

[12] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton,

Adaptive mixtures of local experts, *Neural Computat.*, vol. 3, no. 1, pp. 79–87, 1991.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[14] K. Balog, H. Ramampiaro, N. Takhirov, and K. Nørvåg, Multi-step classification approaches to cumulative citation recommendation, in *Proc. 10th Conference on Open Research Areas in Information Retrieval*, Paris, France, 2013, pp. 121–128.

[15] K. Balog and H. Ramampiaro, Cumulative citation recommendation: Classification vs. ranking, in *Proc. 36th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Dublin, Ireland, 2013, pp. 941–944.

[16] R. Berendsen, E. Meij, D. Odijk, M. de Rijke, and W. Weerkamp, The university of Amsterdam at TREC 2012, in *Proc. 21st Text Retrieval Conf.*, Gaithersburg, MD, USA, 2012.

[17] L. Bonnefoy, V. Bouvier, and P. Bellot, A weakly-supervised detection of entity central documents in a stream, in *Proc. 36th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Dublin, Ireland, 2013, pp. 769–772.

[18] G. G. Gebremeskel, J. Y. He, A. P. de Vries, and J. Lin, Cumulative citation recommendation: A feature-aware comparison of approaches, in *Proc. 25th Int. Workshop on Database and Expert Systems Applications*, Munich, Germany, 2014, pp. 193–197.

[19] J. G. Wang, L. J. Liao, D. D. Song, L. R. Ma, C. Y. Lin, and Y. Rui, Resorting relevance evidences to cumulative citation recommendation for knowledge base acceleration, in *Proc. 16th Int. Conf. on Web-Age Information Management*, Qingdao, China, 2015, pp. 169–180.

[20] J. G. Wang, D. D. Song, Q. F. Wang, Z. W. Zhang, L. Si, L. J. Liao, and C. Y. Lin, An entity class-dependent discriminative mixture model for cumulative citation recommendation, in *Proc. 38th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Santiago, Chile, 2015, pp. 635–644.

[21] F. Chamroukhi, Robust mixture of experts modeling using the t distribution, *Neural Netw.*, vol. 79, pp. 20–36, 2016.

[22] S. R. Waterhouse and A. J. Robinson, Classification using hierarchical mixtures of experts, in *Proc. 1994 IEEE Workshop on Neural Networks for Signal Processing*, Ermioni, Greece, 1994, pp. 177–186.

[23] S. E. Yuksel, J. N. Wilson, and P. D. Gader, Twenty years of mixture of experts, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, 2012.

[24] B. P. Yao, D. B. Walther, D. M. Beck, and F. F. Li, Hierarchical mixture of classification experts uncovers interactions between brain regions, in *Proc 22nd Int. Conf. on Neural Information Processing Systems*, Vancouver, Canada, 2009, pp. 2178–2186.

[25] S. E. Yuksel and P. D. Gader, Variational mixture of experts for classification with applications to landmine detection, in *Proc. 20th Int. Conf. Pattern Recognition*, Istanbul, Turkey, 2010, pp. 2981–2984.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. Ser. B* (*Methodological*), vol. 39, no. 1, pp. 1–38, 1977.

[27] M. Schmidt, MinFunc, http://www.cs.ubc.ca/∼schmidtm/Software/minFunc.html, 2005.

[28] R. Rehurek and P. Sojka, Gensim, https://radimrehurek.com/gensim/, 2010.

[29] N. C. Tu, A Java Implementation of Latent Dirichlet Allocation (LDA), https://sourceforge.net/projects/jgibblda/, 2018.

[30] TREC, KBA Stream Corpus 2013, http://trec-kba.org/kba-stream-corpus-2013.shtml, 2013.

Lerong Ma received the MS degree from Yunnan University, Kunming, China, in 2004. He is currently an associate professor in the School of Mathematics and Computer Science at Yan'an University and the PhD candidate in Beijing Institute of Technology. His research interests include information retrieval, data mining, and natural language processing.

Lejian Liao received the PhD degree from Chinese Academy of Sciences in 1994. He is currently a professor in the School of Computer Science and Technology at Beijing Institute of Technology, Beijing, China. With main reasearch interests in machine learning, natrual language processing, and intelligent network, he has published numerous papers in several areas of computer science.

Dandan Song received the BE and PhD degrees from Tsinghua University, Beijing, China, in 2004 and 2009, respectively. She is currently an associate professor in the School of Computer Science and Technology at Beijing Institute of Technology, Beijing, China. Her research interests include information retrieval, data mining, and bioinformatics.

Jingang Wang received the BS and PhD degrees in computer science from Beijing Institute of Technology (BIT), China, in 2010 and 2016, respectively. Currently, he is a senior algorithm engineer at Alibaba Group. His research interests include information retrieval, knowledge mining, and natural language processing.