TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2018.9010011

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 are valuable resources of human knowledge which have contributed in many applications. However, their manual maintenance makes a big lag between their contents and the up-to-date information of entities. This paper studies Cumulative Citation Recommendation (CCR) - given a target entity in Knowledge Bases, how to effectively detect its worthy-citation documents in large volumes of stream data. In order to build a global relevant model, most previous methods only consider semantic and temporal features of entity-document instances, in which prior knowledge underlying entity-document instances is not exploited sufficiently. To deal with this problem, we present a mixture of experts model by introducing a latent layer to capture relationships between the entity-document instances and their latent class information. An extensive set of experiments have been conducted on TREC-KBA-2013 dataset. The results exhibit the model can achieve a significant performance gain compared to state-of-the-art models in CCR.

Key words: Knowledge base acceleration; cumulative citation recommendation; Mixture of experts; Latent entity-document classes.

1 Introduction

Knowledge Bases (KBs), like Wikipedia, are widely used as a reference tool to search all kinds of information in our daily life. Furthermore, they are playing increasing important roles 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]. Keeping the contents of KBs timely is crucial to these applications. However, most KBs are hard to be up-to-date due to their manual maintenance by human editors. There is a median time lag of one year between the publication date of a news article and the date that the news article is edited into a Wikipedia profile [7]. This time lag would be significantly decreased if documents with highly respect to the target entity in the KBs could be detected automatically as soon as the documents are published online and then recommend them to the editors. This task is studied as Knowledge Base Acceleration-Cumulative Citation Recommendation (KBA-CCR) by the Text REtrieval Conference (TREC). Formally, given a set of Knowledge Base (KB) entities, KBA-CCR is to detect relevant documents from a time-ordered corpus and evaluate their citation-worthy to the target entities.

Due to shortage of training instances for most of target entities, a variety of global relevant supervised models (e.g., classification, learning to rank) have

• Lerong Ma, Lejian Liao and Dandan Song are with Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail:{malerong, Liaolj, 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
• Jingang Wang is with iDST, Alibaba Group, Beijing 100020, China. E-mail: bitwjg@gmail.com
* To whom correspondence should be addressed.
Manuscript received: 2017-03-17; accepted: 2017-08-30

been used in the task and obtained promising performances [8–10]. In most of the models, however, all kinds of features leveraged into the models only capture semantic and temporal information of entities and documents [11]. In fact, in our observation, entities and documents can provide some prior knowledge. For examples, a target entity can offer its categories or its topics, and a document can offer its topics or its source. This prior knowledge embedded in the entity-document instances, called as **class**, would guide the related entity-document selection and thus impact the performance of KBA-CCR. For instance, when processing an entity with a politician category, it would probably have more preferences on a document with politic 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 model to improve the performance of KBA-CCR.

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 the 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 and define the gating functions on the observation data. Our goal is to achieve a mixture of experts that can utilize prior knowledge of entity-document instances in the KBA-CCR task and improve the performance.

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

2 Related Work

2.1 Cumulative Citation Recommendation

TREC launched the KBA-CCR track from 2012 to 2014. Participants treat 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 of them 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, are employed to address this problem [8, 19] with entity-document semantic and temporal features. Nevertheless, characteristics of different entities are lost in the entity-unspecific methods. In [20], latent entity-classes are considered, which have been proven to enhance the performance. Unlike the previous models, the proposed latent entity-document class mixture of experts enhances the model with latent classes information in entity-document pairs in addition to entity-document semantic and temporal information.

2.2 Mixture of Experts Model

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

The mixture of experts has three components. The first one is individual component densities that are 'experts' for making predictions in their own regions. The second one is mixing coefficients known as *gating functions* that determine which components are dominant in which region. The last one is a probabilistic model to combine the experts and the gating functions. Models for the experts in the mixture of experts have been studied in classification tasks by exploiting a variety of different models, such as logistic regression [12], SVM [24] and multinomial [25]. In this paper, we adopt the logistic regression as experts in the paper. Similar to the convention mixture of experts, we make use of softmax function as gating function in our LEDCME model.

3 Mixtures of Experts Models for CCR

This section proposes a novel learning framework for CCR by using a mixture of experts model that combines logistic regression as experts and softmax function as gating functions. The gating function models the latent entity-document classes, and the logistic regression models the relevance of the entity-document instances. We first present a formal definition of the research problem and model it as a classification task, and then propose our Latent Entity-Document Class Mixture

of Experts model (LEDCME) for CCR. Finally, the parameter estimation is given by making use of 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 $\mathcal{E} = \{e_u | u = 1, \cdots, M\}$ and a document collection $\mathcal{D} = \{d_v | v = 1, \cdots, N\}$, our objective is to estimate the relevance of a document d to a given entity e. In other words, we need estimate the conditional probability of relevance $P(r|e, d)$ with respect to an entity-document instance (e, d), where $r \in \{-1, +1\}$ indicate a positive instance if $r = +1$, and a negative instance otherwise. 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 $\mathbf{f}(e, d) = (f_1(e, d), \cdots, f_K(e, d))$, where K indicates the number of entity-document features. The other is for the latent entity-document class information that are represented as a feature vector $\mathbf{g}(e, d) = (g_1(e, d), \cdots, g_L(e, d))$, where L denotes the number of entity-document classes features. The entity-document features and the entity-document class features will be introduced in the **Features** section.

3.2 Entity-Document Class Mixture of Experts Model

The mixture of experts has been applied to the classification tasks [23]. It develops the notion that different components known as "experts" can model the distribution in different regions of input space, and the gating functions decide which expert is prioritized in which region of the input space. As we present the latent entity-document class information, different latent entity-document classes should correspond to different classifiers in order to improve the classification performance. Presumably, the mixture of experts model is suitable for the above cases, so 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 $z \in \{1, 2, \cdots, N_z\}$ as "experts" to capture the latent entity-document class information where N_z is the number of experts, and define

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

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

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

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

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

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

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

According to (3) and (4), the general representation of the expert is given by

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

Finally, we combine the gating function (2) and the experts(5), and obtain the Latent Entity-Document Class Mixture of Experts (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 latent entity-document classes.

3.3 Model Parameter Estimation

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

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

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

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

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

Note that the log-likelihood loss function can exhibit severe over-fitting for the data set $\mathcal{T} = \{(e_u, d_v)|u = 1,\cdots,M; v = 1,\cdots,N\}$ when the data set are linearly separable. One approach that is often used to control the over-fitting phenomenon in such cases is 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:

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

where 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 (6) that will be learned.

As the object function (9) contains latent variables (i.e., the hidden entity-document class z), a typical approach to minimize the object function is to use the Expectation - Maximization (EM) algorithm [26] by iterative E-step and M-step until convergence. Here we have to point out that the standard EM algorithm is to maximum the log-likelihood function, while the loss function (9) is to minimize the negative log-likelihood function, so both the methods are equivalent. In addition, the optimization (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 derivation of the variant EM in detail can be referred in [13].

The E-step can be derived as follows 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 EM algorithm, the variant Q function of the (9) is the following.

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

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

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

and

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

In order to optimize (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 method.

LEDCME has two advantages against the logistic regression. One is that the combination parameters vary across various entity-document classes and hence lead 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 present a Latent Entity Class Mixture of Experts model (LECME) when simplify the model with a single entity class. Thus the LECME model is a special case of the LEDCME model where the major difference is the gating function using the entity class feature vector rather than the entity-document class feature vector. Similar to the LEDCME model, the LECME model is illustrated as:

Given (e, d) denoting an entity-document instance and a target relevant level $r \in \{-1, 1\}$, we introduce a variable $z \in \{1, 2, \cdots, N_z\}$ 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

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

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 $\alpha = (\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kA})$ is a vector of parameters for the gating function. We take the (5) to here

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

Finally, we combine the gating function (14) and the experts (15), and obtain the Latent Entity Class Mixture of Experts model (LECME) written in the following form.

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

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

Similarly, we can obtain the second special case of LEDCME with a single document class, called as Latent Document Class Mixture of Experts model (LDCME), and the representation of LDCME in the form is as follow.

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

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

4 Features

This section proposes two kinds of features used in our LEDCME model. Entity-document features (i.e., $f(e, d)$) are employed in the experts presented in (5). Moreover, LEDCME needs entity-document class features (i.e., $g(e, d)$) to learn the gating functions that correspond to the latent entity-document classes information.

Since our aim is not to develop new entity-document features, we employ the same entity-document feature set presented in the work [8, 19], which have been used effectively, and listed in Table 1.

Table 1 The features of Entity-document pairs

Feature	Description
$N(e_{rel})$	# entity e's related entities found in its profile page
$N(d, e)$	# occurrences of e in document d
$N(d, e_{rel})$	# occurrences of the related entities in document d
$FPOS(d, e)$	First occurrence position of e in d
$FPOS_n(d, e)$	$FPOS(d, e)$ normalized by the document length
$LPOS(d, e)$	Last occurrence position of e in d
$LPOS_n(d, e)$	$LPOS(d, e)$ normalized by the document length
$Spread(d, e)$	$LPOS(d, e) - FPOS(d, e)$
$Spread_n(d, e)$	$Spread(d, e)$ normalized by document length
$Source(d)$	the source of d
$weekday(d)$	weekday of d published
$burst(d)$	burst weights of d

According to the entity-document class information, we consider two groups prior knowledge. One is the prior knowledge of entities, and the other is the prior knowledge of documents. Finally, we combine these two prior knowledges to produce the entity-document class information.

4.1 Prior knowledge of entities

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

Profiled features Every entity in Knowledge Bases, such as Wikipedia and Twitter, has a unique profile page that contains the basic information of this entity, including name, location, biograph and so on. The profile pages of all target entities are acquired from the Wikipedia and Twitter as a profile collection. After preprocessing of the profile collection consisting of removing stop words and stemming, we use the bag-of-word model to represent each target entity as a vector, where term weights are determined by TF-IDF scheme.

Category features Some knowledge bases like Wikipedia curate entities by using hierarchical categories. For instance, **Blair Thoreson** in Wikipedia is labelled with categories including *Member of the north Dakota House of Representatives, 1964 Births, living people, politicians from Fargo, North Dakota*. We append three meta-categories: *person, organization, facility* that cover all the entities in the target entity set. Like profile features, we leverage bag-of-categories to represent the categories of the entity as a vector of features, where category weights are given 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 model topics of documents by making use of bag-of-words and Latent Dirichlet Allocation models. After removing stop words and stemming, we use the gensim [29] package to generate the vector of documents by employing the bag-of-words model, where term weights are determined by TF-IDF scheme. In addition, we use **JGibbLDA** [28], which is a java implementation of Latent Dirichlet Allocation (LDA) using 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 document's topics are 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 news of the Government more reliable than a document from Web Chat. We leverage a 'bag-of-sources' model to represent each document as a feature vector, and term weights are determined in terms of binary occurrence scheme.

5 Experimental

5.1 Dataset

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

Each entity-document instance is assessed as one of 4-point rating levels: (1) **Vital**, timely information of the entity's current state, actions, or situation. This motivates a change to the entity's profile. (2) **Useful**, background information, such as biography, secondary source information. (3) **Neutral**, informative but not citation-worthy information, e.g., tertiary source like Wikipedia articles. and (4) **Garbage**, no information about the target entity can be learned from the document, e.g., spam. The detail annotation of the dataset is listed in Table 2.

Table 2 The detail 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

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.

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 carry out the experiments on a 64-bit machine with Intel Xeon $2.4GH_z$ (L5530), $4MB$ cache and $24GB$ memory. The loss object function (9) involves two hyper-parameters, one is the number of latent entity-document classes 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. The paper utilizes 5-fold cross-validation for selecting the two hyper-parameters on a grid (N_z, λ), where

$$N_z \in \{2, 3, \cdots, 50\}$$

and

$$\lambda \in \{\exp(-50), \exp(-49), \cdots, \exp(0)\}.$$

5.4 Experimental Methodology

Nine variants of LEDCME have been conducted on the TREC-KBA-2013 dataset. In order to further compare with LEDCME, we have also conducted experiments related to Latent Entity Class Mixture of Experts model (LECME) and Latent Document Class Mixture of Experts model (LDCME) by replacing the entity-document class information with only entity class information and only document class information, respectively by setting the other one as only one class.

5.4.1 Latent Entity Class Mixture of Experts model (LECME)

- Profile-based Entity Class Mixture of Experts model (**Profile_LECME**). A variant of LECME utilizes profile-based features as entity class features for the gating function.
- Category-based Entity Class Mixture of Experts model (**Category_LECME**). A variant of LECME utilizes category-based features as entity-class features for the gating function.
- Combine Entity Class Mixture of Experts Model (**Combine _LECME**). A variant of LECME utilizes profile-based and category-based entity features together as entity class features for the gating function. In our experimental setting, we simply join the two types of entity class feature vectors together into an integral feature vector.

5.4.2 Latent Document Class Mixture of Experts model (LDCME)

- Source-based Document Class Mixture of Experts model (**Source_LDCME**). It is a variant of LDCME that uses source-based features as document class features for the gating function.
- TFIDF-based Document Class Mixture of Experts model (**TFIDF_LDCME**). It is a variant of LDCME that uses TFIDF-based features as document class features for the gating function.
- LDA-based Document Class Mixture of Experts model (**LDA_LDCME**). It is a variant of LDCME employs LDA-based features as document class features for the gating function.

5.4.3 Entity-Document Class Mixture of Experts (LEDCME)

(1) Profile Catenating Document Information

- Profile+Source-based Latent Entity-Document Class Mixture of Experts (**Profile+Source_LEDCME**). A variant of LEDCME utilizes profile features of entities catenating source features of documents as the entity-document class features for the gating function.
- Profile+TFIDF-based Latent Entity-Document Class Mixture of Experts (**Profile+TFIDF_LEDCME**). A variant of LEDCME makes use of profile features of entities catenating TF-IDF features of documents as the entity-document class features for the gating function.
- Profile+LDA-based Latent Entity-Document Class Mixture of Experts (**Profile+LDA_LEDCME**). A variant of LEDCME uses profile features of entities catenating LDA features of documents as entity-document class features for the gating function.

(2) Category Catenating Document Information

- Category+Source-based Latent Entity-Document Class Mixture of Experts (**Category+Source_LEDCME**). A variant of LEDCME utilizes category features of entities catenating source features of documents as the entity-document class features for the gating function.
- Category+TFIDF-based Latent Entity-Document Class Mixture of Experts (**Category+TFIDF_LEDCME**). A variant of LEDCME makes use of category features of

entities catenating TFIDF features of documents as the entity-document class features for the gating function.

- Category+LDA-based Latent Entity-Document Class Mixture of Experts (**Category+LDA_LEDCME**). A variant of LEDCME uses category features of entities catenating LDA features of documents as entity-document class features for the gating function.

(3) ProCat catenating Document Information

- ProCat+Source-based Latent Entity-Document Class Mixture of Experts (**ProCat+Source_LEDCME**). A variant of LEDCME utilizes ProCat features of entities catenating source features of documents as the entity-document class features for the gating function, where we append the profile and category features together into an integral features as ProCat features of entities.

- ProCat+TFIDF-based Latent Entity-Document Class Mixture of Experts (**ProCat+TFIDF_LEDCME**). A variant of LEDCME makes use of ProCat features of entities catenating TF-IDF features of documents as the entity-document class features for the gating function, where we append the profile and category features together into an integral features as ProCat features of entities.

- ProCat+LDA-based Latent Entity-Document Class Mixture of Experts (**ProCat+LDA_LEDCME**). A variant of LEDCME uses ProCat features of entities catenating LDA features of documents as entity-document class features for the gating function, where we append 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]. An official baseline in which the annotators manually select a list of keywords of the target entities for filtering vital and useful document.

- **BIT-MSRA** [8]. An entity-unspecific random forests classification model with the first place approach in TREC-KBA-2013 track. This approach extracts 13 types of features between entities and documents, and then learns a global model for all entities using the random forest classification model.

- **UDEL** [9]. An entity-centric query expansion approach that achieves the second performance in TREC-KBA-2013 track. This approach firstly detects related entities from the profile page of a given target entity. Then, the target entity combines the related entities as a new query queries and ranks the relevant documents that have been detected.

- **LR**. The Logistic Regression model on the TREC-KBA-2013 dataset.

5.5 Overall Results

We adopt precision P, recall R and harmonic mean F_1 (harmonic mean between precision and recall) as the evaluation measurements. All the measurements are computed in an entity-insensitive manner. In other words, the measurements are computed based on the test pool of all entity-document pairs regardless of specific entities. Note that low recall and high precision leads to less documents to manually inspect but it may miss important documents. On the other hand, high recall and low precision leads to more documents to review, which may not be feasible if the number of editors is limited. Therefore, we focus on F_1 measurements in the paper.

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

In details, all the variants of LECME outperform all the baselines in both the scenarios. In particular, Combine_LECME achieves significant F1 performance in both the scenarios. This means that the Profile and Category information can enhance each other as the entity class information.

All the variants of LDCME yield better F1 performance in both the scenarios, however, Source_LDCME perform bad in the two scenarios. This intuitively demonstrates that the source of documents is not a crucial factor in determining the importance of documents.

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

Methods	Vital Only			Vital + Useful		
	P	R	F_1	P	R	F_1
Official Baseline	.171	**.942**	.290	.540	**.972**	.694
BIT-MSRA	.214	.790	.337	.589	.974	.734
UDEL	.169	.806	.280	.573	.893	.698
LR	.218	.507	.304	.604	.913	.727
Profile_LECME	.332	.376	.353	.669	.866	.755
Category_LECME	.316	.422	.362	.672	.894	.767
Combine_LECME	.397	.418	.407	.703	.877	**.780**
Source_LDCME	.286	.230	.255	.615	.851	.714
TFIDF_LDCME	.313	.379	.343	.712	.839	.769
LDA_LDCME	.396	.341	.366	**.734**	.828	.778
Profile+Source_LEDCME	.250	.621	.356	.640	.886	.743
Profile+TFIDF_LEDCME	**.405**	.449	.426	.681	.898	.774
Profile+LDA_LEDCME	.331	.584	.422	.639	.870	.737
Category+Source_LEDCME	.281	.478	.354	.628	.909	.744
Category+TFIDF_LEDCME	.403	.454	.427	.674	.903	.771
Category+LDA_LEDCME	.361	.497	.418	.631	.922	.749
ProCat+Source_LECDME	.311	.429	.361	.631	.909	.745
ProCat+TFIDF_LEDCME	.398	.462	**.428**	.685	.882	.772
ProCat+LDA_LEDCME	.404	.416	.410	.646	.892	.749

Among the LEDCME models, the ProCat+TFIDF_LEDCME model achieves the best $F1$ value in the Vital Only scenario, which improves $F1$ by about 47% in contrast to the official baseline. It is also the model with the highest $F1$ value among all the comparative models. In the Vital+Useful scenario, the Profile+TFIDF_LEDCME model achieves the best $F1$ value among the LEDCME models, which increases it by roughly 7% in contrast to the official baseline. In both scenarios, all the variants of LEDCME also outperform the LR model. In comparison to LR, our best model improves $F1$ about 41% in the Vital scenario, and increases $F1$ roughly 7% in the Vital+Useful scenario. These comparisons clearly show the overall effectiveness of our Latent Entity-Document Class Mixture of Experts model.

Moreover, our LEDCME models outperform 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 achieves highest $F1$ values by improving $F1$ values by 5% and 26%, respectively. This indicates that the latent class information in entity-document pairs is more useful than the seperate latent class information in entities and documents in the Vital Only scenario. Similar phenomena appear in other combinations. For example, compared with Category_LECME and TFIDF_LDCME, Category+TFIDF_LEDCME increases F1 by 18% and 25%, respectively. And in contrast with Profile_LECME and TFIDF_LDCME, Profile+TFIDF_LEDCME achieve F1 by 21% and 25%, respectively. These results validate our motivations that (1) incorporating the latent entity-document classes information in mixture of Experts can enhance citation recommendation quality, and (2) Profile and Category features of entities and TFIDF or LDA features of documents can capture the latent entity-document classes information.

Furthermore, all the variants of LEDCME with regard to the source of documents perform worse than the others in the Vital Only scenario. These results meet our expectation in previous discussion about the useless of document sources. And topic-based features of documents including TF-IDF and LDA have far more dimensions than source-based features of documents. However, even if source-based features of documents have only small dimensions (10 in our experiments), Profile+Source_LEDCME, Category+Source_LEDCME and ProCat+Source_LECDME achieve 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 exhibit the catenating entity and document class information strategies are effective; whereas this motivate us to develop further better combination strategies to improve the performance of CCR.

And in the Vital+Useful scenario, the Combine_LECME model achieves 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 Conclusions

The objective of Cumulative Citation Recommendation (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 Latent Entity-Document Class Mixture of Experts (LEDCME) by making use of latent class information in entity-document pairs, with profile and category of the target entities, as well as topics features of documents including TF-IDF and LDA. We have conduct extensive experiments on the TREC-KBA-2013 dataset. The results demonstrate that (i) when introducing the latent entity-document information, the mixture of experts models are effective for CCR, (ii) profile and category of entities and topics and TF-IDF of documents can capture the entity-document class information, and (iii) the catenating entity and document information strategies 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 the performance of CCR.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (Grant Nos. 2016YFB1000902), National Natural Science Foundation of China (Grant Nos. 61472040), and Natural Science Basic Research Plan in Shaanxi Province of China (Grant Nos. 2016JM6082)

References

[1] R. Mihalcea and A. Csomai, Wikify!: linking documents to encyclopedic knowledge, in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Lisbon, Portugal, 2007, pp. 233–242.

[2] Y. Xu, G. J. Jones, and B. Wang, Query dependent pseudo-relevance feedback based on wikipedia, in *Proceedings of the 32nd international ACM SIGIR conference on 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 *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, Queensland, Australia, 2014, pp. 365–374.

[4] Zhang C, Zhou M, Han X, et al. Knowledge graph embedding for hyper-relational data[J]. Tsinghua Science and Technology, 2017, 22(02): 185-197.

[5] H. T. Dang, D. Kelly, and J. J. Lin, Overview of the trec 2007 question answering track, in *Proceedings of The Sixteenth Text REtrieval Conference*, Gaithersburg, Maryland, USA, 2007, Special Publication 500-274.

[6] K. Balog, P. Serdyukov, and A. P. d. Vries, Overview of the trec 2010 entity track, in Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, 2010.

[7] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Ré, and I. Soboroff, Building an entity-centric stream filtering test collection for trec 2012, DTIC Document, Tech. Rep., 2012.

[8] J. Wang, D. Song, C.-Y. Lin, and L. Liao, Bit and msra at trec kba ccr track 2013, in *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013*, Gaithersburg, Maryland, USA, 2013.

[9] X. Liu, J. Darko, and H. Fang, A related entity based approach for knowledge base acceleration, in *Proceedings of The Twenty-Second Text REtrieval Conference, TREC*, Gaithersburg, Maryland, USA 2013.

[10] J. Frank, S. J. Bauer, M. Kleiman-Weiner, D. A. Roberts, N. Triouraneni, C. Zhang, and C. Rè, Evaluating stream filtering for entity profile updates for trec 2013, in *Proceedings of The Twenty-Second Text REtrieval Conference, TREC*, Gaithersburg, Maryland, USA 2013.

[11] K. Balog and H. Ramampiaro, Cumulative citation recommendation: Classification vs. ranking, in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, 2013, pp. 941–944.

[12] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, Adaptive mixtures of local experts, *Neural computation*, 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 *OAIR*. ACM, 2013, pp. 121–128. [Online]. Available: http://dl.acm.org/citation.cfm?id=2491748.2491775

[15] K. Balog and H. Ramampiaro, Cumulative citation recommendation: classification vs. ranking, in *The 36th International ACM SIGIR conference on research and development in Information Retrieval*, Dublin, Ireland, 2013, pp. 941–944.

[16] R. Berendsen, E. Meij, D. Odijk, M. d. Rijke, and W. Weerkamp, The university of amsterdam at trec 2012, in *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, 2012*.

[17] L. Bonnefoy, V. Bouvier, and P. Bellot, A weakly-supervised detection of entity central documents in a stream, in *the 36th International ACM SIGIR conference on research and development in Information Retrieval*, Dublin, Ireland, 2013, pp. 769–772.

[18] G. G. Gebremeskel, J. He, A. P. d. Vries, and J. Lin, Cumulative citation recommendation: A feature-aware comparison of approaches, in *Database and Expert Systems Applications (DEXA)*. IEEE, 2014, pp. 193–197.

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

[20] J. Wang, D. Song, Q. Wang, Z. Zhang, L. Si, L. Liao, and C.-Y. Lin, An entity class-dependent discriminative mixture model for cumulative citation recommendation, in *Proceedings of the 38th International ACM SIGIR Conference on 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 Networks*, vol. 79, pp. 20–36, 2016.

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

[23] S. E. Yuksel, J. N. Wilson, and P. D. Gader, Twenty years of mixture of experts, *IEEE transactions on neural networks and learning systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[24] B. Yao, D. Walther, D. Beck, and L. Fei-Fei, Hierarchical mixture of classification experts uncovers interactions between brain regions, *Advances in Neural Information Processing Systems*, 2009, pp. 2178–2186.

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

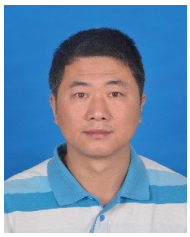
[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[27] http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html

[28] http://jdibblda.sourceforge.net

[29] https://radimrehurek.com/gensim/

[30] http://trec-kba.org/kba-stream-corpus-2013.html



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

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

Dandan Song received a B.E degree and Ph.D. degree from the Department of Computer Science and Technology, Tshinghua University, Beijing, China, in 2004 and 2009, respectively. She is currently an associated professor in the School of Computer Science and Technology, 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.