A Hybrid Discriminative Mixture Model for Cumulative Citation Recommendation

Lerong Ma, Dandan Song, Lejian Liao, Jingang Wang

Abstract—This paper explores Cumulative Citation Recommendation (CCR) for Knowledge Base Acceleration (KBA). The CCR task aims to detect potential citations of a set of target entities with priorities from a volume of temporally-ordered stream corpus. Previous approaches for CCR that build an individual relevance model for each entity fail to deal with unseen entities without annotation. A compromised solution is to build a global entity-unspecific model for all entities without respect to the relationship information among entities, which cannot guarantee to achieve satisfactory result for each entity. Moreover, most previous methods can not adequately exploit prior knowledge embedded in entities or documents due to considering all kinds of features indifferently. In this paper, we propose a novel entity and document class-dependent discriminative mixture model by introducing one intermediate layer to model the correlation between entity-document pairs and hybrid latent entity-document classes. The model can better adjust to different types of entities and documents, and achieve better performance when dealing with a broad range of entity and document classes. An extensive set of experiments has been conducted on two offical datasets, and the experimental results demonstrate that the proposed model can achieve the state-of-the-art performance.

Index Terms—Cumulative Citation Recommendation, Knowledge Base Acceleration, hybrid latent entity-document classes,Mixture Model.

1 Introduction

In recent years, we have witnessed a proliferation of Knowledge Bases (KBs) such as Wikipedia and Freebase[1]. Nowadays, KBs are widely used as quick reference tools to find all kinds of information in our daily lives, and they have shown great power in several application scenarios such as query answering, entity search, and entity linking. Here, KBs are defined as collaborative online encyclopaedia edited and maintained by their users. KBs are usually organized as entities to store the facts about the real world, and an entity can be a person, facility, organization, or concept such as "Barack Obama", "White House", "Democratic Party", or "Car".

Currently, the maintenance of a KB mainly relies on human editors. However, with the explosion of information, large-scale KBs are hard to be kept up-to-date solely by human editors. Taking English Wikipedia for example, there are approximately 4.7 million entities but merely 132,938 active editors[2]. The less popular entities cannot be updated in time because they are not spotlighted. As reported in [1], the median time delay between a cited document's publishing and its citation in Wikipedia is almost one year. An outdated KB severely limits the effectiveness of applications depending on it. This gap could be bridged if relevant documents of KB entries can be automatically detected as soon as they emerge online and then be recommended to the editors with corresponding relevance levels. This task is studied as Cumulative Citation Recommendation (CCR) in research communities, which was launched by TREC since 2012. Formally, given a KB entity, CCR is a task to filter highly relevant documents from a volume of chronological stream corpus and evaluate their citation-worthiness to the target entity.

Most previous approaches (e.g., [2], [3], [4], [5]) for CCR are highly supervised and require sufficient training data to build an individual relevance model for each entity. These approaches are infeasible when dealing with a large-scale KB (e.g., Wikepedia), since the annotation work is labor intensive to obtain enough training data for each entity. A compromised solution is to build a global entity-unspecific discriminative model and optimize it to achieve an overall local optimal performance for all entities in training data (e.g., [6], [7]). However, these models ignore the distinctions between different entities and learn a set of fixed model parameters for all entities, which leads to unsatisfactory performance when dealing with a diverse entity set. For instance, it is not fair to apply the same discriminative model for *Geoffery Hinton* and *Appleton Museum of Art*. The former entity is a computer scientist, while the latter one is a museum. Therefore, with respect to *Geoffery Hinton*, the combination weights for *Barbara Liskov* (a female computer scientist) is more helpful than that of *Appleton Museum of Art*. Nevertheless, the global model treats them equally without considering this prior knowledge of the distinctions among entities. For this findings, Wang et al. [8] propose an entity class-dependent discriminative mixture model for CCR by introducing a latent entity class layer to model the correlations between entities and latent entity classes, and the model achieved better performance in KBA-CCR-2013 dataset.

- *Lerong Ma, Dandan Song, Lejian Liao and Jingang Wang are with Beijing Engineering Research Center of High Volumn Language Information Processing and Cloud Computing Application, Beijing Lab of Intelligence Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, 100081 Beijing, China E-mail: {malerong_bit, sdd, bitwjg,liaolj}@bit.edu.cn*
- *Lerong Ma is also School of Mathematics and Computer Science, Yan'an University, 716000 Yan'an, China*
- *Dandan Song is corresponding author*

1. https://www.freebase.com/
2. http://meta.wikimedia.org/wiki/List_of_Wikipedias#1_000_000.2B_articles

In fact, the task of CCR considers the relevances between entities and documents. Therefore, we should consider entity prior knowledge and document prior knowledge together that a target entity can offer its categories or its topics, and a document can offer its topics. For instance, when handling a document of the "music" topic, it is reasonable to put less weights on a politician entity, because this document is not likely to be related to a politician but more often related to musicians or musical bands. This motivates us to leverage the prior knowledge of entities and documents together into the model to improve the performance of KBA-CCR.

Inspired by the above motivations, we propose a hybird entity and document class dependent discriminative mixture model (HEDCDMM) for CCR by utilizing the underlying entity and document class information together. The model introduces a hybrid latent entity and document class layer to define a joint distribution over the entity-document pair and the hybrid latent entity and document classes conditioned on the observations. When the model only consider latent entity class information given the observation, the model adjusts to the entity class-dependent discriminative mixture model [8], where the latent entity class layer defines a joint distribution over the entity-document pairs and the latent entity classes. In the same way, when the model only consider latent document class information, the model is adaptive to the document class-dependent discriminative mixture model, where the latent document class layer defines a joint distribution over the entity-document pairs and latent document classes. The aim is to achieve relevance estimation through learning a mixture model which is expected to outperform the global model, while maintaining the capability to reveal the hidden correlations for the combination of entities and documents vs. the hybrid latent entity and document classes. Compared with a classification model learned for each entity in the KB, the proposed models use latent entity or document class information that exploits the prior knowledge of entities and documents. The proposed models are not only global models but also leverages latent entity and latent document class information together, which would be applied to construction of a large-scale KB.

The model can be viewed as a hierarchical combination of a mixing component and a discriminative component, so two types of features are required: entity-document features for the discriminative component, and latent class features for the entities and documents features for the mixing component. For the discriminative component, we use semantic features between entities and documents, and a bursty feature of entities detected from the internal stream corpus. For the mixing component, we explore two types of entity-class features including profile-based features and category-based features. Profile-based features are constructed from the entity's profile in KB, while category-based features rely on the existing category labels for the entity in KB. Moreover, we develop two kinds of document-class features based on topics of documents to capture the prior knowledge of documents. One is TF-IDF model, and the other is Latent Dirichlet Allocation (LDA) topic model.

To the best of our knowledge, this is the first research work that focuses on modeling correlations for the combination of entities and documents vs. hybrid latent entity and document classes in discriminative model for CCR. Our model is capable of tackling less popular entities with small training data and unseen entities that do not exist in the training set, which is indispensable in a practical CCR system. Empirical studies have been conducted on TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 datasets to prove the effectiveness and robustness of the proposed mixture models. Experimental results demonstrate that our models achieves the state-of-the-art performance on TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 datasets.

The main contributions of this paper can be summarized as follows:

- We consider the prior knowledge of entities and documents together, and differentiate entity and document class features from other features, such as semantic feature between entities and documents.
- We propose a hybrid entity and document class-dependent discriminative mixture model (HEDCDMM) for CCR by utilizing the latent entity class information and the latent document class information together. The entity-class dependent mixture model for CCR [8] and the document-class dependent mixture model are two special models for HEDCDMM.
- We conducted extensive experiments on TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 datasets. Experimental results demonstrate that the proposed model achieves higher or competitive performances on the datasets compared with other reference approaches, and considerable gains are obtained for long-tail entities as well as obtain better results among unseen entities in the testing data.

2 Related Work

Although CCR was first proposed in TREC-KBA tracks, the similar research problem had been studied in several topics of information retrieval.

Topic/Event Detection and Tracking

Topic Detection and Tracking (TDT) is a track hosted by TREC from 1997 to 2004 [9]. A similar research topic in recent years is event detection. Both TDT and event detection are concerned with the development of techniques for finding and following events in broadcast news or social media. The techniques adopted for TDT and event detection can be broadly classified into two categories: (1) clustering documents based on the semantic distance between them [10], or (2) grouping the frequent words together to represent events [11]. In [11], a finite automaton model is proposed to detect events in stream by modeling events as state transitions. This method has been validated widely by lots of other studies [12], [13], [14]. We also adopt this model to detect KB entities' bursts in the stream corpus and then extract bursty features for them. Different from above works, we model entities' occurrences to capture bursty activities instead of words' occurrences. Another difference between CCR and TDT is that CCR needs to make fine-grained citation-worthiness distinctions between relevant documents further.

Cumulative Citation Recommendation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2893328, IEEE Transactions on Knowledge and Data Engineering

TREC has launched the KBA-CCR track since 2012. Participants treat CCR as either a ranking problem [2], [5], [15] or a classification problem [5], [6], [16]. Classification and Learning to Rank methods were compared and evaluated [2], [17], and both of them can achieve performance well with a powerful feature set. Several supervised learning techniques, such as SVMs [18], [19], language models [20], [21], [22], Markov Random Fields [23], and Random Forests [6], [15], [16] were utilized. Meanwhile, a variety of relevance scoring methods were tried, including standard Lucene scoring [24], and custom ranking functions based on entity co-occurrences [3]. A time-aware evaluation paradigm is developed to study time-dependent characteristics of CCR [25].

However, some highly supervised methods [18], [19] 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 [4], [6], [26]. Nevertheless, characteristics of different entities are lost in the entity-unspecific methods. Some other researchers employ transfer learning techniques to learn across entities by using entity-unspecific meta-features [7], or utilize a semi-supervised approach to profile an entity by leveraging its related entities and weighting them with the training data [27]. These methods have demonstrated that the correlations between entities are useful for CCR. Nevertheless, all these methods are empirically designed and the performance can be improved further. Moreover, query expansion is often employed because the name of the target entity is too sparse to be a good query. Other name variants and contextual information of terms or related entities from Wikipedia or from the document stream [23], [24] are used to enrich the semantic features of entities. In addition to semantic features, temporal features have been proved especially helpful in CCR [2], [6], [16].

Recently, there is one deep learning approach to handle the CCR task [28], termed as DeepJoED. DeepJoED presents a Joint Deep Neural Network Model of Entities and Documents to identify highly related documents for given entities with several layer neuron networks, and trains the networks in an end-to-end fashion. However, due to computing resource limitation, DeepJoED only conducts esperiments on TREC-KBA-2012 dataset.

Mixture Model

Mixture model has been proved effective to address the problem of data insufficiency in several information retrieval tasks, including expert search [29], federated search [30], collaborative filtering [31] and image retrieval [32]. By introducing latent layers to learn flexible combination weights for different feature vectors, mixture model can always outperform simple discriminative models with fixed combination weights. Hence, we propose a hybrid entity and document class-dependent discriminative mixture model to deal with the entities with small training data, which will be described in next section.

3 Approach Overview

We propose an approach to address CCR task. Figure 1 shows an overview of our approach that consists of several key tasks. In the first step we extract features from entities and documents, and the second step is classification given the entity-document pairs using the hybrid discriminative mixture model. When the model classifies documents from the text stream corpus as relevant given the target entities within a knowledge base, editors for the knowledge base update the corresponding contents of the target entities.

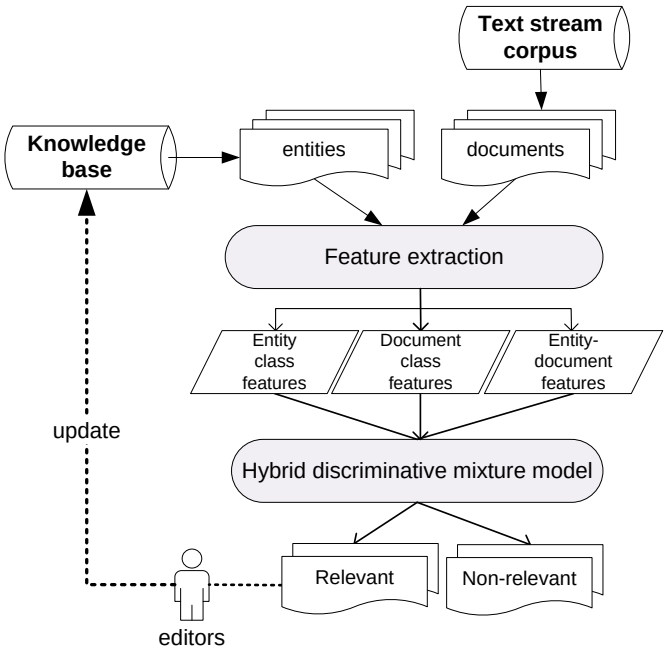


Fig. 1. Approach overview of CCR task. In the first step we extract features from entities and documents, and the second step is classification given the entity-document pairs using the hybrid discriminative mixture model.

4 Discriminative Mixture Model for CCR

This section proposes a novel learning framework that combines a logistic regression model with each entity's distribution across hidden entity classes, each document's distribution over underlying document classes, or each entity-document's distribution over hybrid entity and document classes to form a final discriminative mixture model. First we formulate the research problem and model it as a classification task, and then propose four discriminative models: a global discriminative model, a hybrid entity-document class-dependent discriminative mixture model and its two special casses: an entity class-dependent discriminative mixture model, a document class-dependent discriminative mixture model.

4.1 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. Many probabilistic classification techniques in the literature generally fall into two categories: generative models and discriminative models. Discriminative models have attractive theoretical properties [33] and generally perform better than their generative counterparts in the field of information retrieval [34], [35]. Therefore, we adopt discriminative probabilistic models in this paper.

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 pair (e, d). Each entity-document pair (e, d) is 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. Moreover, to model the hidden entity classes information, each entity can be represented as an entity-class feature vector $\mathbf{g(e)} = (g_1(e), \cdots, g_L(e))$, where L indicates the number of entity-class features. Furthermore, to capture the latent document classes information, each document is represented as document-class feature vector $\mathbf{g(d)} = (g_1(d), \cdots, g_C(d))$, where C is the number of document-class features. The entity-document features, entity-class features and document-class features will be introduced in Section 5 later.

4.2 Global Discriminative Model

This paper utilizes logistic regression, a traditional discriminative model, to estimate the conditional probability $P(r|e, d)$ given an entity-document pair observation (e, d), in which $r(r \in \{1, -1\})$ is a binary label to indicate the relevance of the entity-document pair (e, d). The value of r is 1 if the document d is relevant to the entity e, -1 otherwise. Formally, the parametric form of $P(r = 1|e, d)$ can be expressed as follows in terms of logistic functions over a linear combination of features,

$$P(r = 1|e, d) = \delta(\sum_{i=1}^{K} \omega_i f_i(e, d) + b) \tag{1}$$

where $\delta(x) = 1/(1 + \exp(-x))$ is the standard logistic function, b is a constant, and ω_i is a combination weight for the ith entry of the entity-document feature vector. For representation simplicity, we let $\omega_0 = b$ and $f_0(e, d) = 1$, then Eq. 1 is rewritten as follow

$$P(r = 1|e, d) = \delta(\sum_{i=0}^{K} \omega_i f_i(e, d)). \tag{2}$$

We follow this representation convention in the following of the paper. As for the irrelevant class, we have

$$P(r = -1|e, d) = 1 - P(r = 1|e, d) = \delta(-\sum_{i=0}^{K} \omega_i f_i(e, d)) \tag{3}$$

It is worth noting that for different values of r, the only difference in $P(r|e, d)$ is the sign within the logistic function. Therefore, we adopt the general representation of

$$P(r|e, d) = \delta(r \sum_{i=0}^{K} \omega_i f_i(e, d)) \tag{4}$$

in the following sections. The conditional probability of relevance $P(r|e, d)$ represents the extent to which the document d is relevant to the entity e. The entity-document pairs are then classified as positive instances or negative instances according to the value of $P(r = 1|e, q)$. Since the learned weights are identical for all the entity-document pairs and regardless of specific entities, this model is also denoted as global discriminative model (GDM) in this paper.

Several other approaches for CCR [6], [16] can be deemed as global discriminative models adopting different classification functions such as decision trees and Support Vector Machine (SVM).

4.3 Hybrid Entity and Document Class Dependent Mixture Model

Given a entity-document pair, we can consider not only the latent entity classes but also the underling document classes. Therefore, we propose a hybrid entity and document class-dependent discriminative mixture model (**HEDCDMM**) to capture the entity class information and the document class information together. Specifically, we define two latent random variables z and x indicating the entity class and the document class, respectively. The choice of z and x is determined by the target entity and document in the entity-document pair (e, d). The joint probability of relevance r, the latent random variable z and the latent random variable x is represented as

$$P(r, z, x|e, d; \alpha, \beta, \omega) = P(z, x|e, d; \alpha, \beta)P(r|e, d, z, x; \omega) \tag{5}$$

where $P(z, x|e, d; \alpha, \beta)$ is the mixing coefficient indicating the joint probability of latent entity class z and hidden document class x given the entity-document pair (e, d), and α and β are the corresponding parameters. $P(r|e, d, z, x; \omega)$ denotes the mixture component which takes a logistic functions for $r = 1$ (or $r = -1$). $\omega = \{\omega_{zxi}, i = 1, \cdots, K\}$ is the set of parameters where ω_{zxi} is the weight for the ith entry of feature vector to the given training instance (e, d) under the latent entity class z and the latent document class x.

Since the generation processes of entities in a KB and doucment from a stream corpus are different, we assume that the random variable z is independent of the random variable x. In other words, the hidden entity class information are independent of the latent document class information. We hence rewrite the above joint probability as follow

$$P(r, z, x|e, d; \alpha, \beta, \omega) = P(z|e; \alpha)P(x|d; \beta)\delta\left(r \sum_{i=0}^{K} \omega_{zxi} f_i(e, d)\right) \tag{6}$$

Figure 2 illustrates the representation of HEDCDMM graphic model, where α, β and ω are parameters learned by the model.

If $P(z|e; \alpha)$ follows the multinomial distribution, the model cannot easily generalize the combination weights to unseen entities beyond the training set since each parameter in multinomial distribution specifically corresponding to a training entity. To address this problem, we adopt a soft-max function to model $P(z|e; \alpha)$ as follow

$$P(z|e; \alpha) = \frac{1}{Z_e} \exp(\sum_{j=1}^{L} \alpha_{zj} g_j(e)) \tag{7}$$

where α_{zj} is the weight parameter associated with the jth entity class feature in the latent entity class z, and Z_e is the normalization factor that scales the exponential function to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2893328, IEEE Transactions on Knowledge and Data Engineering

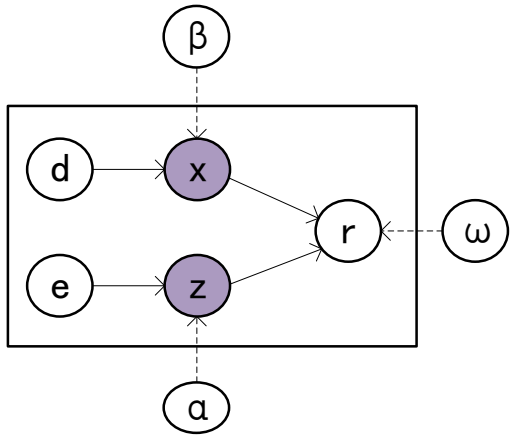


Fig. 2. The representation of HEDCDMM graphic model.α, β and ω are parameters learned by the model.

be a proper probability distribution. In this representation, each entity e is denoted by a bag of entity-class features $(g_1(e), \cdots, g_L(e))$ where L is the number of entity-class features.

Similarly, we define $P(x|d;\beta)$ as follow

$$P(x|d;\beta) = \frac{1}{X_d}\exp(\sum_{j=1}^{C}\beta_{xj}g_j(d)) \tag{8}$$

where β_{xj} is the weight parameter associated with the jth document-class feature vector in the latent document class x, and X_d is the normalization factor that scales the exponential function to be a proper probability distribution. Each document d is represented by a bag of document class features $(g_1(d), \cdots, g_C(d))$ where C is the length of document-class features.

We take Eq. 7 and Eq. 8 into Eq. 6, and marginalize out the latent variables z and x. **HEDCDMM** can be represented in the following.

$$\begin{aligned} P(r|e,d;\alpha,\beta,\omega) = \\ \frac{1}{Z_e}\frac{1}{X_d}\sum_{z=1}^{N_z}\sum_{x=1}^{N_x}\exp\left(\sum_{j=1}^{L}\alpha_{zj}g_j(e)\right) \\ \exp\left(\sum_{j=1}^{C}\beta_{xj}g_j(d)\right)\delta\left(r\sum_{i=0}^{K}\omega_{zxi}f_i(e,d)\right). \end{aligned} \tag{9}$$

where N_z is the number of latent entity classes, and N_x is the number of latent document classes.

Note that here we follow the convention in Eq. 4, in which $w_{zx0} = b$ and $f_0(e,d) = 1$.

We suppose entity-document pairs in training set are represented as $\mathcal{T} = \{(e_u, d_v)\}$, and $\mathcal{R} = \{r_{uv}\}$ denotes the corresponding relevance judgment (i.e., $+1$ or -1) of (e_u, d_v), where $u = 1, \cdots, M$ and $v = 1, \cdots, N$. Assume training instances in $\mathcal{T}$ are independently generated, the likelihood of the training data is written as follows.

$$\begin{aligned} P(\mathcal{R}|\mathcal{T}) = \prod_{u=1}^{M}\prod_{v=1}^{N}P(r_{uv}|e_u,d_v) = \\ \prod_{u=1}^{M}\prod_{v=1}^{N}\left(\frac{1}{Z_{e_u}}\frac{1}{X_{d_v}}\sum_{z=1}^{N_z}\sum_{x=1}^{N_x}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u))\right. \\ \left.\exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{xi}f_i(e_u,d_v))\right) \end{aligned} \tag{10}$$

4.4 Parameter Estimation

The parameters (i.e. ω, α and β) in Eq. 10 can be estimated by maximizing the following data log-likelihood functions

$$\begin{aligned} \mathcal{L}_h(\omega,\alpha,\beta) = \\ \sum_{u=1}^{M}\sum_{v=1}^{N}\log\left(\frac{1}{Z_{e_u}}\frac{1}{X_{d_v}}\sum_{z=1}^{N_z}\sum_{x=1}^{N_x}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u))\right. \\ \left.\exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zxi}f_i(e_u,d_v))\right) \end{aligned} \tag{11}$$

where M is the number of the entities and N is the number of the documents in training set. $g_j(e_u)$ denotes the jth feature for the uth entity, $g_j(d_v)$ indicates the jth feature for the corresponding vth document and r_{uv} denotes the relevance judgment for the pair (e_u, d_v). A typical approach to maximize the equation Eq. 11 above is to use Expectation-Maximization (EM) algorithm [36].

For Eq. 11, E-step can be derived as follows by computing the joint posterior probability of z and x given an entity-document pair (e_u, d_v), α, β and ω. Let

$$\begin{aligned} Norm(e_u,d_v) = \sum_{z=1}^{N_z}\sum_{x=1}^{N_x}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u)) \\ \exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zxi}f_i(e_u,d_v)), \end{aligned}$$

and we have

$$\begin{aligned} P(z,x|e_u,d_v) = \frac{1}{Norm(e_u,d_v)}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u)) \\ \exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zxi}f_i(e_u,d_v)). \end{aligned} \tag{12}$$

In terms of Eq. 12 we can derive the conditional posterior probability of $P(z|e_u,d_v)$ and $P(x|e_u,d_v)$ as follows.

$$\begin{aligned} P(z|e_u,d_v) = \frac{1}{Norm(e_u,d_v)}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u)) \\ \sum_{x=1}^{N_x}\exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zxi}f_i(e_u,d_v)) \end{aligned} \tag{13}$$

and

$$P(x|e_u,d_v)=\frac{1}{Norm(e_u,d_v)}\exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v)) \sum_{z=1}^{N_z}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u))\delta(r_{uv}\sum_{i=0}^{K}\omega_{zxi}f_i(e_u,d_v)). \tag{14}$$

For M-step of Eq. 11, we can obtain the following parameters update rules with the auxiliary Q function using the above posterior probabilities of Eq. 12, Eq. 13 and Eq. 14.

$$\omega_{zx}^{*}=\arg\max_{\omega_{zx}}\sum_{u=1}^{M}\sum_{v=1}^{N}P(z,x|e_u,d_v) \log\left(\delta(r_{uv}\sum_{i=0}^{K}\omega_{zxi}f_i(e_u,d_v))\right) \tag{15}$$

$$\alpha_{z}^{*}=\arg\max_{\alpha_z}\sum_{u=1}^{M}\left(\sum_{v=1}^{N}P(z|e_u,d_v)\right) \log\left(\frac{1}{Z_{e_u}}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e_u))\right) \tag{16}$$

$$\beta_{x}^{*}=\arg\max_{\beta_x}\sum_{v=1}^{N}\left(\sum_{u=1}^{M}P(x|e_u,d_v)\right) \log\left(\frac{1}{X_{d_v}}\exp(\sum_{j=1}^{C}\beta_{xj}g_j(d_v))\right) \tag{17}$$

The corresponding optimization problems of the M-step in the above equations can be optimized by any gradient descent method. In this paper we employ the `minFunc` toolkit[3], a collection of Matlab functions for solving optimization problems using Quasi-Newton strategy. When the corresponding optimization problems above converge to a local optima, the estimated parameters can be plugged back into the corresponding model to compute the probability of relevance for entity-document pairs. Since EM is only guaranteed to converge to local optima given different starting points, we try several starting points and choose the model that leads to the greatest log-likelihood.

4.5 Two special cases of HEDCDMM

Eq. 5 defines two latent random variables z and x that model not only latent entity class inforation but also latent document class information. When we only consider latent entity class information, i.e., the latent entity random variable z, HEDCMM reduces an entity class-depenedent discrimative mixture model(**ECDMM**).

More specifically, the random variable z is utilized to indicate which entity class the combination weights $\omega_z = (\omega_{z1},\cdots,\omega_{zK})$ are drawn from. The choice of z depends on the target entity e in the entity-document pair (e,d). The joint probability of relevance r and the latent variable z is represented as

$$P(r,z|e,d;\alpha,\omega)=P(z|e;\alpha)P(r|e,d,z;\omega) \tag{18}$$

3. http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html

where $P(z|e;\alpha)$ is the mixing coefficient, representing the probability of choosing hidden entity class z given entity e, and α is the corresponding parameters. $P(r|e,d,z;\omega)$ denotes the mixture component which takes a logistic functions for $r=1$ (or $r=-1$). $\omega=(\omega_{z1},\cdots,\omega_{zK})$ is the vector of combination parameters where ω_{zi} is the weight for the ith feature vector entry for the given training instance (e,d) under the hidden class z. By marginalizing out the latent variable z, the corresponding mixture model can be written as

$$P(r|e,d;\alpha,\omega)=\sum_{z=1}^{N_z}P(z|e;\alpha)\delta\left(r\sum_{i=0}^{K}\omega_{zi}f_i(e,d)\right) \tag{19}$$

where N_z is the number of latent entity classes. In the same way, we define $P(z|e;\alpha)$ as a soft-max function $\frac{1}{Z_e}\exp(\sum_{j=1}^{L}\alpha_{zj}g_j(e))$, α_{zj} is the weight parameter associated with the jth entity class feature in the latent entity class z and Z_e is the normalization factor that scales the exponential function to be a proper probability distribution. By plugging the soft-max function into Eq. 19, we can get **ECDMM** model as follow

$$P(r|e,d;\alpha,\omega)=\frac{1}{Z_e}\sum_{z=1}^{N_z}\exp\left(\sum_{j=1}^{L}\alpha_{zj}g_j(e)\right) \delta\left(r\sum_{i=0}^{K}\omega_{zi}f_i(e,d)\right). \tag{20}$$

Similarly, when we only consider latent doucment class information, Eq. 5 changes to documentan class-depenedent discrimative mixture model(**DCDMM**). The representation of **DCDMM** is written as follow:

$$P(r|e,d;\beta,\omega)= \frac{1}{X_d}\sum_{x=1}^{N_x}\exp\left(\sum_{j=1}^{C}\beta_{xj}g_j(d)\right)\delta\left(r\sum_{i=0}^{K}\omega_{xi}f_i(e,d)\right) \tag{21}$$

where N_x is the number of latent document classes, ω_{xi} is the weight for the ith entry in the document-class feature vector under the hidden random variable x, and X_d is the normalization factor that scales the exponential function to be a proper probability distribution.

4.6 Discussion

HEDCDMM, ECDMM and DCDMM can take the following advantages against the GDM: (1) the combination weights are able to change across entities and hence lead to a gain of flexibility. (2) it offers probabilistic semantics for the latent entity class information and thus each entity can be associated with multiple latent entity classes. (3) it also support probabilistic semantics for the underlying document class information and each document is related to multiple latent document classes. (4) it offer probabilistic semantics for the hybrid entity and document class information, thus each entity-document can be associated with multiple latent entity and document classes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2893328, IEEE Transactions on Knowledge and Data Engineering

5 FEATURES

5.1 Entity-Document Features

Entity-document features (i.e., $\mathbf{f}(e, d)$) are composed of semantic and temporal features. The semantic and temporal features are listed in Table 1, which have been proved effective in CCR [4], [6]. Semantic features can model semantic characteristics of entity-document pairs. Temporal features model the dynamic characteristics of entities. If the number of documents related to an entity rises suddenly in a short time period there must have something important happen about the entity. Accordingly, this short time period can be detected as one bursty for the entity. Therefore, we make an assumption documents published in a bursty period of an entity are more likely related to the entity. Signals such as Wikipedia page views have been shown to be effective [8]. However, such signals are spare or not available for long-tail entities that lack any profiles. Therefore, we employ intrinsic document stream as data source to detect the bursty periods of target entities using internal burst detection [8] in the paper.

TABLE 1
Semantic and Temporal features

Feature	Description
$N(e_{rel})$	# of entity e's related entities found in its profile page
$N(d, e)$	# of occurrences of e in document d
$N(d, e_{rel})$	# of 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
$Sim_{cos}(d, s_i(e))$	Cosine similarity between d and the ith section of e's profile
$Sim_{jac}(d, s_i(e))$	Jaccard similarity between d and the ith section of e's profile
$Sim_{cos}(d, c_i)$	Cosine similarity between d and the ith citation of e in the KB
$Sim_{jac}(d, c_i)$	Jaccard similarity between d and the ith citation of e in the KB
$Burst(e, d)$	Burst value of the document d in the burst period of the entity e

5.2 Entity-Class Features

In addition to entity-document features, entity-class features are required to learn the mixing coefficients. Here we present two types of prior knowledge to design entity-class features used in the ECDMM and HEDCDMM.

5.2.1 Profile-based features

Each entity in KBs is uniquely identified by its profile page, which contains the basic information of this entity, such as name, address and experiences. We crawl the profile pages of all the entities as a profile collection. After removing stop words, we represent each entity as a feature vector with the bag-of-words model, where term weights are determined by the TF-IDF scheme.

5.2.2 Category-based features

Some KBs like Wikipedia organize entities with hierarchical categories. For example, `Geoffrey Hinton` in Wikipedia, is labeled with categories such as `Canadian computer scientists`, `Artificial intelligence researchers`, and `Fellows of AAAI`. Besides these labeled categories, we take the parent categories of the labeled categories into consideration to deal with the circumstance in Figure 3. The two alike entities can not be correlated if we only consider labeled categories.

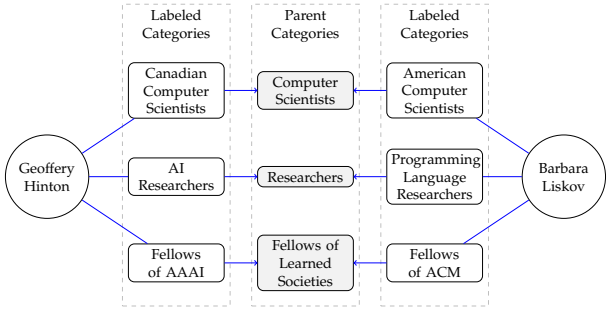


Fig. 3. Two entities without common labeled categories but with shared parent categories.

Similar to profile-based feature vector, we leverage a "bag-of-categories" model to represent each entity as a category-based feature vector. Given an entity without category information, we manually assign a meta-category for it according to its profile. We supplement three meta-categories: **person**, **facility** and **organization**, which can cover all the entities in our dataset. The category-based feature vector of entity e is denoted as $g^c(e) = (c_1(e), \cdots, c_N(e))$, where N is the total number of categories. $c_i(e)$ equals to 1 if e is labeled with category c_i, otherwise $c_i(e)$ is 0.

Therefore, given a target entity set $\mathcal{E}$, we can generate two feature vectors for each $e \in \mathcal{E}$: profile-based vector $\mathbf{g}^p(e)$ and category-based vector $\mathbf{g}^c(e)$ respectively.

5.3 Document-Class Features

Here we propose one type of prior knowledge to represent document-class features used in the DCDMM and HEDCDMM. The prior knowledge to model underlying class of a document is its intrinsic topics. Documents with one or more explicit topics are more likely related to entities having similar topics. We capture underlying topics in documents of the corpus with word co-occurrences and Latent Dirichlet Allocation (LDA) topic model. For the word co-occurrences, after removing stop words, we represented each document as a feature vector using the bag-of-words model, where word weights are computed by TF-IDF schema. For the LDA topic model, in order to obtain topic distribution of documents in stream corpus we employ the JGibbLDA [4]

4. https://sourceforge.net/projects/jgibblda/

package, a Java Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference. We keep 10 thousand words as the dictionary and use 500 topics under run JGibbLDA.

Therefore, given a set of document in the corpus D, we can yield two feature vectors for each document $d \in D$: TFIDF-based vector and LDA-based vector respectively.

6 Experiments

6.1 Dataset

We conduct our experiments on TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 datasets[5], three standard test beds provided by TREC. They come from 10 sources: news, mainstream news, social, weblog, linking, arxiv, classified, reviews, forum and memetracker. Each data set is composed of a target entity set and a document collection called stream corpus.

Entity Set

The entity set of TREC-KBA-2012 consists of 29 Wikipedia enities including 27 persons and 2 organizations. These entities are described by semi-structured articles in Wikipedia. For the TREC-KBA-2013 dataset, the target entity set includes 121 Wikipedia entities and 20 Twitter entities, more specifically, 98 people, 19 organizations, and 24 facilities from 14 inter-related communities such as small towns like `Danville, KY` and academic communities like `Turing award winners`. For the TREC-KBA-2014, the set of target entities consists of 70 entities based on people, organizations, and facilities. Entities have an profile in the form of Wikipedia, Twitter, or None indicating that no entity profile is built as a description of the entity. Table 2 provides a breakdown of entity profile types for TREC-KBA official datasets.

TABLE 2
Distribution of entity profile types for TREC-KBA-XXXX official datasets

Entity type	2012	2013	2014
Wikipedia	29	121	28
Twitter	0	21	0
None	0	0	42

Stream Corpus

The stream corpus for TREC-KBA-2012 includes documents crawled from news, social media, and Linking, which cover the period from October 2011 to April 2012. The corpus is divided as training and testing instances, with documents from October to December 2012 as training instances, and the remainder for testing. The paper follows this setup.

The TREC-KBA-2013 dataset is a temporally-ordered stream corpus, containing approximately 1 billion documents crawled from October 2011 to the end of February 2013. Each document is associated with a timestamp indicating its time of crawling. The corpus have been split with documents from October 2011 to February 2012 as training instances and the remainder for evaluation. We adopt the same training/test range setting in our experiments.

5. http://trec-kba.org

The TREC-KBA-2014 corpus expanded the TREC-KBA-2013 corpus, covering the time period from October 2011 to May 2013. The whole corpus has 1.2 billion documents. Each entity in TREC-KBA-2014 dataset has its own training period and testing period, that is, documents whose timestamps are in the training period of the entity are treated as training instances, and the documents whose timestamps are in the testing period of the entity are considered as testing instances.

Annotation The relevance of entity-document pairs are labeled following a four-point scale relevance setting, including *vital*, *useful*, *neural* and *garbage*. The definitions are listed in Table 3.

TABLE 3
Four-point scale relevance estimation in TREC-KBA-2013 and TREC-KBA-2014.

Vital	timely info about the entity's current state, actions, or situation. This would motivate a change to an already up-to-date KB article.
Useful	possibly citable but not timely, e.g., background biography, secondary source information.
Neutral	informative but not citable, e.g., tertiary source like Wikipedia article itself.
Garbage	no information about the target entity could be learned from the document, e.g., spam.

The details of the annotations for TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 are demonstrated in Table 4. We note that the number of training instances of TREC-KBA-2014 is less than those of TREC-KBA-2013, yet the size of stream corpus of TREC-KBA-2014 is bigger than those of TREC-KBA-2013.

6.2 Evaluation Scenarios

According to different granularity settings, TREC-KBA evaluates models in two classification scenarios respectively.

Vital Only

Only *vital* entity-document pairs are treated as positive instances, and the others are treated 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 are negative ones. Note that the core task of CCR is for the *Vital Only* scenario, and the *Vital+useful* scenario is just for reference.

6.3 Experimental Methodology

Twelve variants of discriminative models are implemented to dedicatedly evaluate the effectiveness of our proposed mixture models. They are conducted on the two datasets and compared in the two scenarios.

6.3.1 Basic Approaches

- Global Discriminative Model (**GDM**). As presented in Subsection 4.2, this model learns a set of fixed weights for all entity-documents pairs.
- Naïve Class-Dependent Discriminative Mixture Model (**Naïve_CDMM**). This approach simply uses entity-document features as the mixing component (i.e., $\mathbf{g}(e) := \mathbf{f}(e, d)$ in Eq. 20) of ECDMM.

TABLE 4
The number of training and testing instances (entity-document pairs) annotated for TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 respectively.

Datasets	Training					Testing				
	Vital	Useful	Neutral	Garbage	total	Vital	Useful	Neutral	Garbage	total
TREC-KBA-2012	3525	6500	1757	9382	21264	5252	8426	2470	20439	36587
TREC-KBA-2013	938	4320	481	3196	8935	3202	11608	505	503	15818
TREC-KBA-2014	2278	2583	1234	2325	8420	10447	19006	7140	22797	59390

6.3.2 Entity Class Dependent Mixture Models (ECDMM)

- Profile-based Entity Class Dependent Discriminative Mixture Model (**Profile_ECDMM**). It is a variant of ECDMM that utilizes profile-based features as entity-class features for the mixing component.
- Category-based Entity Class-Dependent Discriminative Mixture Model (**Category_ECDMM**). It is a variant of ECDMM that utilizes category-based features as entity-class features for the mixing component.
- Combined Entity Class Dependent Discriminative Mixture Model (**Combine_ECDMM**). It is a variant of ECDMM that combine profile-based and category-based features together as entity-class features for the mixing component. In our current implementation setting, we union the two feature vectors together into an integral feature vector.

6.3.3 Document Class Dependent Mixture Models (DCDMM)

- TFIDF-based Document Class-Dependent Discriminative Mixture Model (**TFIDF_DCDMM**). It is a variant of DCDMM that uses TFIDF-based features as document-class features for the mixing component of DCDMM.
- LDA-based Document Class-Dependent Discriminative Mixture Model (**LDA_DCDMM**). It is a variant of DCDMM that employs category-based features as entity-class features for the mixing component of DCDMM.

6.3.4 Hybrid Entity and Document Class Dependent Mixture Models (HEDCDMM)

- Hybrid Entity Profile and Document TFIDF based Class Dependent Discriminative Mixture Model (**Pro2TFIDF_HEDCDMM**). It is a variant of HEDCDMM that uses Profile-base features of entity classes and TFIDF-based features of document classes for the mixing component.
- Hybrid Entity Profile and Document LDA based Class Dependent Discriminative Mixture Model (**Pro2LDA_HEDCDMM**). This approach utilizes profile-based features of entity class and LDA features of document classes for the mixing component of HEDCDMM.
- Hybrid Entity Category and Document TFIDF based Class Dependent Discriminative Mixture Model (**Cat2TFIDF_HEDCDMM**). This approach utilizes category-based features of entity classes as well as TFIDF-based features of document classes for the mixing component of HEDCDMM.
- Hybrid Entity Category and Document LDA based Class Dependent Discriminative Mixture Model (**Cat2LDA_HEDCDMM**). It is a variant of HEDCDMM that utilizes category-based features of entity classes together LDA-based features of document classes for the mixing component.
- Hybrid Entity Profile and Category as well as Document TFIDF based Class Dependent Discriminative Mixture Model (**ProCat2TFIDF_HEDCDMM**). It is a variant of HEDCDMM that uses the profile and category features of the entity classes as well as TFIDF features of the document classes as the mixing component.
- Hybrid Entity Profile and Category as well as Document LDA based Class Dependent Discriminative Mixture Model (**ProCat2LDA_HEDCDMM**). This method employs the profile and category features of the entity classes as well as LDA features of the documents classes as the mixing component of HEDCDMM.

For further references, we include the best official run **HLTCOE** [18] from TREC 2012 and **PSVM-2L_300** [19] for KBA-CCR-2012, **BIT-MSRA** [6] and **UDEL** [27] for KBA-CCR-2013 as well as **Official Baseline 2014** [37] for KBA-CCR-2014.

6.4 Experimental Setting

The ECDMM, DCDMM and HEDCDMM models involve hyper-parameters including the number of hidden classes of entity, the number of hidden classes of document, and the number of hybrid hidden classes of entity and document, respectively. Since Akaike Information Criteria (AIC) used in [8] may lead to overfit when we consider the number of hybrid classes of entity and document, we utilizes 5-fold cross-validation for selecting hyper-parameters of the presented model in this paper. For ECDMM and DCDMM, we choose the one that has the best harmonic mean F_1 with the number of hidden entity classes HE and the number of hidden document classes HD respectively, where $HE, HD \in \{2, 3, 4 \cdots, 50\}$. For HEDCDMM, we choose the pair that has the best harmonic mean F_1 with the hybrid number of hidden entity and document classes (HE, HD), where $HE, HD \in \{2, 3, 4 \cdots, 50\}$. Subsequently, we learn the models using the whole training data. Finally, we leverage the learned models to predict instances of the testing set.

6.5 Results

We adopt precision P, recall R and harmonic mean F_1 (harmonic mean between precision and recall) as the eval-

uation metrics. All the metrics are computed in an entity-insensitive manner. That is, the metrics 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 candidate documents to be manually inspected by editors, but it may miss important documents. On the other hand, high recall and low precision produces more documents, which may not be feasible when the number of editors is limited. Therefore, we focus on the F_1 metric in this paper. The results for TREC-KBA-2012, TREC-KBA-2013 and TREC-KBA-2014 are reported in Table 5, Table 6 and Table 7 respectively. In order to compute statistic significance between different approaches, we determine per-entity performance scores using the best threshold with which an approach obtain the best F_1 scores, and then calculate statistic significance between different approaches using the per-entity performnace score lists.

6.5.1 Overall results

In the *Vital Only* scenario, it is shown that ProCat2LDA_HEDCDMM obtains the best F_1 scores on TREC-KBA-2012, Cat2LDA_HEDCDMM performs the best $F1$ on TREC-KBA-2013, and ProCat2LDA_HEDCDMM achieves the best F_1 performance on TREC-KBA-2014. These demonstrate the effectiveness of our hybrid discriminative model. The two official baselines achieve the best recall among all methods on the TREC-KBA-2013 and TREC-KBA-2014 respectively, which is not surprising since the TREC-KBA-2013 official baseline is a method to detect as many relevant documents as possible by manually selecting reliable aliases of an entity in advance, and the TREC-KBA-2014 official baseline queries the relevant documents by using the canonical name of entities as the surface names of the target entities [37].

Compared with GDM which do not incorporate either entity class information or document class information, all the mixture models employing entity-class features, document-class features, or hybrid entity and document class features explicitly achieve better classification performance in both scenarios on the thtree datasets. This reveals that the mixture model is an effective strategy to enhance the straightforward discriminative model. In contrast to GDM, ProCat2LDA_HEDCDMM, Cat2LDA_HEDCDMM and ProCat2LDA_HEDCDMM increase F_1 for more than 40%, 53% and 69% in the *Vital Only* scenario on the three datasets, respectively.

Naïve_CDMM is not robust in the two scenarios on the three datasets. Although it outperforms GDM in the *vital + useful* scenario, it cannot beat GDM in the *Vital Only* scenario on the TREC-KBA-2013 dataset. However, it is slightly better in harmonic mean F_1 than GDM on the TREC-KBA-2014 dataset. This is possibly caused by its implicitly employment of entity-document class features. In contrast to Naïve_CDMM, all the mixture approaches perform better in the two scenario on the three datasets. This shows that prior knowledge of entities and documents can help improve cumulative citation recommendation performance.

All the mixture methods outperform the reference approaches in the *Vital Only* scenario on TREC-KBA-2012 and TREC-KBA-2013. However, we notice that PSVM-2L_300 obtains the best F_1 scores in the *Vital+Useful* scenario on TREC-KBA-2012, which is attributed to PSVM-2L_300 using additional preference information between different level instances [19].

6.5.2 Results of ECDMM

Among the entity class dependent discriminative models, both profile_ECDMM and category_ECDMM remarkably outperform naïve_CDMM, GDM as well as other baselines in the two scenarios on the three datasets except for PSVM-2L_300. This reveals that profile-based features and category-based features are effective in modeling hidden entity classes. Category-based features are more promising than profile-based features, which is reasonable because the category labels in KBs are given by human editors. Even though Combine_ECDMM combines profile-based features and category-based features in a straightforward manner, it achieves better performance. The F_1 of Combine_ECDMM is increased by 22%, 12% and 41% against Naïve_CDMM in the *Vital Only* scenarios on the three datasets, respectively. The p-values between Combine_ECDMM and Naïve_CDMM for the three datasets in the *Vital Only* scenario are 0.032, 0.047, 0.026, respectively, which are all less than 0.05 showing statistically significant.

In the *Vital Only* scenario, Combine_ECDMM achieves considerable higher F_1 scores than Profile_ECDMM and Category_ECDMM on TREC-KBA-2013, and p-values between Combine_ECDMM and Profile_ECDMM and Category_ECDMM are 0.034 and 0.006, respectively. However, Combine_ECDMM obtains a marginal higher F_1 scores than Profile_ECDMM and Category_ECDMM on TREC-KBA-2012 and TREC-KBA-2014, and p-values between Combine_ECDMM and these are greater 0.05. This may be due to the fact that TREC-KBA-2013 have many more Wikipedia entities (refer to Table 2) than other two datasets, which include much human editors's knowledge of target entities.

6.5.3 Results of DCDMM

Among the document class discriminative models, both TFIDF_DCDMM and LDA_DCDMM exceed GDM and naïve_CDMM remarkably in the two scenarios on the three datasets. This shows that topic features of documents are effective in modeling hidden document class information. On TREC-KBA-2012 and TREC-KBA-2013, both TFIDF_DCDMM and LDA_DCDMM do not perform better than profile and categorey-based methods, and in comparison to Combine_ECDMM, LDA_DCDMM are decreased by roughly 7% and 10% respectively, and p-values between Combine_ECDMM and LDA_DCDMM are 0.003 and $5.01E-5$. This show Wikipedia profile class information is more promising than document class information because Wikipedia entities on these two datasets constitute the majority of target entities. LDA_DCDMM is slightly higher than Combine_ECDMMM in the *Vital Only* scenario on TREC-KBA-2014, and p-value between LDA_DCDMM and Combine_ECDMMM is 0.184 which shows the differences in F_1 performance is not statistically significant. This results may be caused by the fact that long-tail entities constitute a large fraction of the target entities (42 entities, i.e., 60%), which lack profile.

TABLE 5
Overall classification results of evaluated models on TREC-KBA-2012 dataset.

Methods	Vital Only			Vital + Useful		
	P	R	F_1	P	R	F_1
HLTCOE	.319	.536	.359	**.695**	.451	.492
PSVM-2L_300	.328	.588	.333	.666	.823	**.717**
GDM	.259	.739	.325	.264	**.845**	.403
Naïve_CDMM	.243	**.866**	.338	.293	.810	.431
Profile_ECDMM	.267	.711	.388	.350	.824	.492
Category_ECDMM	.360	.467	.406	.353	.822	.495
Combine_ECDMM	.326	.555	.411	.476	.715	.572
TFIDF_DCDMM	.265	.710	.386	.348	.794	.484
LDA_DCDMM	.306	.519	.385	.322	.792	.457
Pro2TFIDF_HEDCDMM	.347	.526	.418	.527	.758	.622
Pro2LDA_HEDCDMM	.381	.531	.443	.537	.764	.630
Cat2TFIDF_HEDCDMM	.350	.579	.436	.503	.808	.620
Cat2LDA_HEDCDMM	**.408**	.498	.449	.547	.745	.632
ProCat2TFIDF_HEDCDMM	.373	.572	.452	.509	.805	.624
ProCat2LDA_HEDCDMM	.397	.532	**.454**	.511	.792	.627

TABLE 6
Overall classification results of evaluated models on 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
GDM	.218	.507	.304	.604	.913	.727
Naïve_CDMM	.223	.400	.286	.627	.912	.744
Profile_ECDMM	.332	.376	.353	.669	.866	.755
Category_ECDMM	.316	.422	.362	.672	.894	.767
Combine_ECDMM	.397	.418	.407	.703	.877	.780
TFIDF_DCDMM	.313	.379	.343	.712	.839	.769
LDA_DCDMM	.396	.341	.366	.734	.828	.778
Pro2TFIDF_HEDCDMM	.420	.497	.455	.719	.813	.763
Pro2LDA_HEDCDMM	.403	.507	.449	.730	.878	**.797**
Cat2TFIDF_HEDCDMM	.397	.506	.445	.698	.849	.766
Cat2LDA_HEDCDMM	**.425**	.517	**.466**	.741	.839	.787
ProCat2TFIDF_HEDCDMM	.353	.588	.441	.705	.845	.773
ProCat2LDA_HEDCDMM	.370	.571	.449	**.756**	.802	.779

TABLE 7
Overall classification results of evaluated models on TREC-KBA-2014 dataset.

Methods	Vital Only			Vital + Useful		
	P	R	F_1	P	R	F_1
Offical Baseline	.078	**.989**	.145	.406	**.993**	.578
GDM	.137	.864	.236	.651	.949	.772
Naïve_CDMM	.159	.665	.258	.656	.954	.777
Profile_ECDMM	.238	.577	.338	.657	.968	.783
Category_ECDMM	.219	.727	.338	**.659**	.983	**.789**
Combine_ECDMM	.255	.633	.364	.658	.969	.784
TFIDF_DCDMM	**.413**	.297	.346	.653	.990	.787
LDA_DCDMM	.324	.425	.367	.652	.991	.787
Pro2TFIDF_HEDCDMM	.295	.488	.368	.651	.965	.778
Pro2LDA_HEDCDMM	.271	.601	.374	.651	.962	.776
Cat2TFIDF_HEDCDMM	.359	.402	.379	.650	.959	.775
Cat2LDA_HEDCDMM	.279	.642	.389	.653	.964	.778
ProCat2TFIDF_HEDCDMM	.344	.460	.393	.651	.965	.778
ProCat2LDA_HEDCDMM	.321	.528	**.399**	.651	.955	.774

6.5.4 Results of HEDCDMM

Compared with all the variants of ECDMM and DCDMM, all variants of Hybrid Entity and Document Class Dependent Mixture models perform better in the *vital Only* scenario on the three datasets. On TREC-KBA-2012, compared with top-ranked Combine_ECDMM and TFIDF_DCDMM in the Table 5, ProCat2LDA_HEDCDMM increases F_1 for 11% and 15%, respectively, in the *Vital Only* scenario. Moreover, p-values between ProCat2LDA_HEDCDMM and top Combine_ECDMM and TFIDF_DCDMM are 0.014 and 0.001. Likely, on TREC-KBA-2013, in contrast to top-ranked Combine_ECDMM and LDA_DCDMM in Table 6, Cat2LDA_HEDCDMM improve $F1$ for 14% and 27% respectively in the *Vital Only* scenario, and p-values between those are $1.9E-09$ and $5.5E-05$, which show the improvements in F_1 are statistically significant. These results validate our initial motivation that incorporating the prior knowledge of entity and document class information together can improve KBA-CCR performance to identify the vital document from the text streams.

From the Table 7, in comparison to top-ranked Combine_ECDMM and LDA_DCDMMM, ProCat2LDA_HEDCDMM increases in F_1 for 9% and 8% on TREC-KBA-2014 dataset in the *Vital Only* scenario, and p-values between those are 0.11 and 0.18, which show the differences in F_1 are not statistically significant. This may be due to the fact that there are a fairly large number of long-tail entities in TECK-KBA-2014, which lack profile. In order to gain underlying insights into the results of TREC-KBA-2014, we select a top-ranked method for different methodology setting (i.e., GDM, Combine_ECDMM, LDA_DCDMM and ProCat2LDA_HEDCDMM), and segment the results by entity popularity using the types of entity profile in the Table 2. More specifically, we determine the best threshold for each approach, and compute the performances of Wikipedia segment and None segment using the best threshold. The segment results are presented in Table 8.

TABLE 8
Different segment results of top performance methods in the *Vital Only* scenario on TREC-KBA-2014

Methods	Wikipedia segment			None segment		
	P	R	F_1	P	R	F_1
GDM	.119	.861	.210	.167	.816	.278
Combine_ECDMM	.196	.558	.291	.273	.642	.383
LDA_DCDMM	.289	.344	.314	.349	.495	.409
ProCat2LDA_HEDCDMM	.264	.493	**.344**	.384	.559	**.455**

From the results reported in Table 8, the performances of Combine_ECDMM, LDA_DCDMM and ProCat2LDA_HEDCDMM in None segment achieve higher or competitive precision, recall and F_1 than counterparts on Wikipedia segment. One important factor for these results is that these are most likely long-tail entities with very few candidate documents to consider. Moreover, LDA_DCDMM achieves a comparable performance as compared to GDM and Combine_ECDMM on the None segment, which shows that hidden document class information is effective to detect vital document for long-tail entities that lack profiles. More importantly, ProCat2LDA_HEDCDMM outperforms GDM, Combine_ECDMM and LDA_DCDMM in F_1 considerably on None segment. In particular, the improvements in F_1 in this segment are statistically significant, and p-values between ProCat2LDA_HEDCDMM and GDM, Combine_ECDMM and LDA_DCDMM are 0.001, 0.003 and 0.013. This finding is important and interesting because it verifies the effectiveness of the provided approach for long-tail entities. In fact, filtering vital documents from text stream for target long-tail entities is very important task for a knowledge base construction. Thus, we can adopt ProCat2LDA_HEDCDMM to filter vital documents for long-tail entities.

Among all variants of HEDCDMM in the *Vital Only* scenario on different datasets, there is no clear winner from the results shown as in Table 5, Table 6 and Table 7. These may be due to the facts that the sets of target entities in different datasets are different, distributions of entity types refered to Table 2 are various, and the number of annotation instances of three datasets are different, which are reported in Table 4. However, Cat2LDA_HEDCDMM obtains the highest precision scores among methods On TREC-KBA-2012 and TREC-KBA-2013, as well as achieves top F1 performances, which meet our expectation because the categories of target entities in Wikipedia have much human knowledge. Therefore, when we consider the update of target entities from Wikipedia, we can adopt Cat2LDA_HEDCDMM.

6.5.5 Discussion

From results of all the approaches on TREC-KBA-2012 in Table 5, TREC-KBA-2013 in Table 6 and TREC-KBA-2014 dataset in Table 7, TREC-KBA-2014 CCR task is harder than other two CCR task because the entities in TREC-KBA-2014 dataset are primarily long-tail items, which most items lack profiles, yet only have descriptive pages from web or no any descriptive contents. From the above analysis of the results, we can conclude that (1) HEDCDMM outperforms ECDMM and DCDMM in most scenarios on the three datasets, which validates our motivation that hybrid entity and document class information can enhance CCR performance. Specific model selection suggestions are provided below. (2) When information of entities, such as category, profile and training documents, is fully available, which corresponds to TREC-KBA-2012, we can use ProCat2LDA_HEDCDMM to deal with the CCR task. (3) When training data is not sufficient, Cat2LDA_HEDCDMM is suggested which obtains the highest precision and F1 scores among methods on TREC-KBA-2013. It meets our expectation because the categories of target entities in Wikipedia have much human knowledge. (4) When information of entities is not fully available, which corresponds to TREC-KBA-2014, we can adopt ProCat2LDA_HEDCDMM to filter vital documents for long-tail entities.

Moreover, when we probe the content of missing and wrong documents for a target entity to ProCat2LDA_HEDCDMM in detailed examples, we found some common failures of our approach, they are mainly that: (1) When a document with the same class of a target entity reports an event regarding something else that just mentions the target entity, the document would be a wrong mention by our approach. (2) When a document indirectly

mentions the event of the target entity, the document will be missed by our approach.

6.6 Performance on Unseen Entities

This section evaluates the generalization ability of our proposed models to handle unseen entities. A robust model should be able to handle not only training entities but also unseen entities. We evaluate the performance of our models on the unseen entity set composed of these 10 entities listed in Table 9. Due to fewer positive instances for unseen entities, it is improper to adopt precision, recall and F_1 for evaluation because they possibly become 0, in which case these measurements cannot reflect the performance suitably. We choose macro-averaged accuracy as the evaluation metrics. The results are reported in Table 10.

In *vital only* scenarios, the top two classification results are achieved by LDA_DCDMM and Cat2LDA_HEDCDMM which outperform other variants of our mixture models. This shows document topics based on LDA are robust to model hidden class information. However, Category_ECDMM obtain unsatisfactory performance in both scenarios on unseen entities. Especially, Category_ECDMM performs bad in *Vital+Useful* scenarios, that leads to the performances of Cat2TFIDF_HEDCDMM and Cat2LDA_HEDCDMM lower than TFIDF_DCDMM and LDA_DCDMM respectively. A possible explanation for unsatisfactory performance of Category_ECDMM is that the category information of unseen entities are not covered well in the training set, especially the Twitter entities. For the Twitter entities, there are too little category information to model their hidden classes accurately.

In *vital+Useful* scenario, all the variants of our mixture models can achieve better classification performance than GDM and the other three baselines. The results verify the flexibility of our mixture model as expectation. Our mixture models are not only good at handling existing entities in training set, but also capable of dealing with unseen entities. The flexibility of property of our mixture models are essential for a practical CCR system that is usually required to process amounts of novel entities never existing before. Therefore, the proposed models would be applied to the construction of a large-scale KB.

7 Conclusions and Future Work

The objective of Cumulative Citation Recommendation (CCR) is to detect citation-worthy documents for a set of KB entities from a chronological stream corpus. To address the problem of training data insufficiency for less popular entities, we propose the entity class-dependent discriminative mixture model (ECDMM) , document class-dependent discriminative mixture model (DCDMM) and hibrid entity and document class-dependent discriminative mixture model (HEDCDMM) by employing entity and document class information on the basis of global discriminative model (GDM). Entity-document features, entity-class features and document-class features are proposed for corresponding models. Experimental results demonstrate that when incorporating entity class information and document class information, the mixture models can improve classification performance considerably.

For future work, we wish to explore more useful entity-class features and document-class features as well as apply more proper combination strategies to improve CCR performance.

Acknowledgments

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, 61751217, 61866038), and PH.D start project of Yan'an University of China (Grant Nos.YDBK2018-09).

References

[1] 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," in *TREC*. NIST, 2012.
[2] K. Balog and H. Ramampiaro, "Cumulative citation recommendation: classification vs. ranking," in *SIGIR*. ACM, 2013, pp. 941–944.
[3] A. D. O. Gross and H. Toivonen, "Term association analysis for named entity filtering," in *TREC*. NIST, 2012.
[4] J. Wang, L. Liao, D. Song, L. Ma, C. Lin, and Y. Rui, "Resorting relevance evidences to cumulative citation recommendation for knowledge base acceleration," in *WAIM*, 2015.
[5] K. Balog, H. Ramampiaro, N. Takhirov, and K. Nørvåg, "Multi-step classification approaches to cumulative citation recommendation," in *OAIR*. ACM, 2013, pp. 121–128.
[6] J. Wang, D. Song, L. Liao, and C. Lin, "Bit and msra at trec kba ccr track 2013," in *TREC*. NIST, 2013.
[7] M. Zhou and K. C. Chang, "Entity-centric document filtering: boosting feature mapping through meta-features," in *CIKM*. ACM, 2013, pp. 119–128.
[8] J. Wang, D. Song, Q. Wang, Z. Zhang, L. Si, L. Liao, and C. Lin, "An entity class-dependent discriminative mixture model for cumulative citation recommendation," in *SIGIR*, 2015, pp. 635–644.
[9] J. Allan, "Introduction to topic detection and tracking," in *Topic Detection and Tracking*, ser. The Information Retrieval Series. Springer US, 2002, vol. 12, pp. 1–16.
[10] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *SIGIR*. ACM, 1998, pp. 28–36.
[11] J. M. Kleinberg, "Bursty and hierarchical structure in streams," in *KDD*. ACM, 2002, pp. 91–101.
[12] Q. He, K. Chang, E. Lim, and J. Zhang, "Bursty feature representation for clustering text streams." in *SDM*. SIAM, 2007, pp. 491–496.
[13] Q. He, K. Chang, and E. Lim, "Using burstiness to improve clustering of topics in news streams," in *ICDM*. IEEE, 2007, pp. 493–498.
[14] X. Zhao, R. Chen, K. Fan, H. Yan, and X. Li, "A novel burst-based text representation model for scalable event detection," in *ACL*. ACL, 2012, pp. 43–47.
[15] R. Berendsen, E. Meij, D. Odijk, M. de Rijke, and W. Weerkamp, "The university of amsterdam at trec 2012," in *TREC*. NIST, 2012.
[16] L. Bonnefoy, V. Bouvier, and P. Bellot, "A weakly-supervised detection of entity central documents in a stream," in *SIGIR*. ACM, 2013, pp. 769–772.
[17] G. G. Gebremeskel, J. He, A. P. de Vries, and J. J. Lin, "Cumulative citation recommendation: A feature-aware comparison of approaches," in *Database and Expert Systems Applications (DEXA)*. IEEE, 2014, pp. 193–197.
[18] B. Kjersten and P. McNamee, "The hltcoe approach to the trec 2012 kba track," in *TREC*. NIST, 2012.
[19] L. Ma, D. Song, L. Liao, and J. Wang, "PSVM: a preference-enhanced SVM model using preference data for classification," *SCIENCE CHINA Information Sciences*, vol. 60, no. 12, pp. 122 103:1–122 103:14, 2017.
[20] S. Araújo, G. G. Gebremeskel, J. He, C. Boscarino, and A. P. de Vries, "Cwi at trec 2012, kba track and session track," in *TREC*. NIST, 2012.
[21] L. Dietz and J. Dalton, "Umass at trec 2013 knowledge base acceleration track," in *TREC*. NIST, 2013.

TABLE 9
The statistics of test instances for 10 unseen entities.

Entity	KB	vital	useful	neutral/garbage	total
The Ritz Apartment (Ocala,Florida)	Wiki	4	1	5	10
Keri Hehn	Wiki	3	0	0	3
Chiara Nappi	Wiki	2	3	55	60
Chuck Pankow	Wiki	7	0	10	17
John H. Lang	Wiki	2	0	1	3
Joshua Boschee	Wiki	191	23	5	219
MissMarcel	Twitter	52	13	3	68
evvnt	Twitter	1	3	40	44
GandBcoffee	Twitter	0	2	2	4
BartowMcDonald	Twitter	1	18	9	28

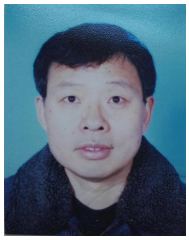
TABLE 10
The averages of accuracies over 10 unseen entities.

Methods	accu@(vital)	accu@(vital + useful)
Official Baseline	.175	.532
BIT-MSRA	.445	.614
UDEL	.259	.579
GDM	.552	.565
naïve_CDMM	.587	.608
Profile_ECDMM	.623	.647
Category_ECDMM	.565	.431
Combine_ECDMM	.580	.582
TFIDF_DCDMM	.615	**.719**
LDA_DCDMM	.688	.595
Pro2TFIDF_HEDCDMM	.506	.601
Pro2LDA_HEDCDMM	.517	.628
Cat2TFIDF_HEDCDMM	.681	.558
Cat2LDA_HEDCDMM	**.697**	.577
ProCat2TFIDF_HEDCDMM	.657	.547
ProCat2LDA_HEDCDMM	.642	.609

[22] S. Kawahara, K. Seki, and K. Uehara, "Detecting vital documents in massive data streams," *OJWT*, vol. 2, no. 1, pp. 16–26, 2015.
[23] J. Dalton and L. Dietz, "Bi-directional linkability from wikipedia to documents and back again: Umass at trec 2012 knowledge base acceleration track," in *TREC*. NIST, 2012.
[24] C. Tompkins, Z. Witter, and S. G. Small, "Sawus: Siena's automatic wikipedia update system," in *TREC*. NIST, 2012.
[25] L. Dietz and J. Dalton, "Time-aware evaluation of cumulative citation recommendation systems," in *SIGIR 2013 Workshop on Time-aware Information Access (TAIA2013)*, 2013.
[26] R. Reinanda, E. Meij, and M. de Rijke, "Document filtering for long-tail entities," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM*, 2016, pp. 771–780.
[27] X. Liu, J. Darko, and H. Fang, "A related entity based approach for knowledge base acceleration," in *TREC*. NIST, 2013.
[28] L. Ma, D. Song, L. Liao, and Y. Ni, *Cluster Comput (2017)*, https://doi.org/10.1007/s10586-017-1273-x.
[29] Y. Fang, L. Si, and A. Mathur, "Discriminative probabilistic models for expert search in heterogeneous information sources," *Information Retrieval*, vol. 14, no. 2, pp. 158–177, 2011.
[30] D. Hong and L. Si, "Mixture model with multiple centralized retrieval algorithms for result merging in federated search," in *SIGIR*. ACM, 2012, pp. 821–830.
[31] R. Jin, L. Si, and C. Zhai, "A study of mixture models for collaborative filtering," *Information Retrieval*, vol. 9, no. 3, pp. 357–382, 2006.
[32] Q. Wang, L. Si, and D. Zhang, "A discriminative data-dependent mixture-model approach for multiple instance learning in image classification," in *ECCV*, 2012, pp. 660–673.
[33] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 841–848.
[34] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, 2007.
[35] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *SIGIR*. ACM, 1999, pp. 42–49.
[36] 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.
[37] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, E. M. Voorhees, and I. Soboroff, "Evaluating stream filtering for entity profile updates in trec 2012, 2013, and 2014," in *TREC*. NIST, 2014.

Lerong Ma received his PH.D. degree from Beijing Institute of technology(BIT), Beijing, China, in 2018. He is currently an associated professor in the School of Mathematics and Computer Science, Yan'an University, Yan'an, China. His research interests include knowlege bases, information retrieval, data mining and natural language processing.

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. Her research interests include information retrieval, data mining and bioinformatics



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. 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.

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.