LDTM: A Latent Document Type Model for Cumulative Citation Recommendation

Jingang Wang[1]*, Dandan Song[1]†, Zhiwei Zhang[2], Lejian Liao[1], Luo Si[2], Chin-Yew Lin[3]
[1]School of Computer Science, Beijing Institute of Technology
[2]Dept. of Computer Science, Purdue University
[3]Knowledge Mining Group, Microsoft Research
{bitwjg, sdd, liaolj}@bit.edu.cn
{zhan1187, lsi}@purdue.edu
cyl@microsoft.com

Abstract

This paper studies Cumulative Citation Recommendation (CCR) - given an entity in Knowledge Bases, how to effectively detect its potential citations from volume text streams. Most previous approaches treated all kinds of features indifferently to build a global relevance model, in which the prior knowledge embedded in documents cannot be exploited adequately. To address this problem, we propose a latent document type discriminative model by introducing a latent layer to capture the correlations between documents and their underlying types. The model can better adjust to different types of documents and yield flexible performance when dealing with a broad range of document types. An extensive set of experiments has been conducted on TREC-KBA-2013 dataset, and the results demonstrate that this model can yield a significant performance gain in recommendation quality as compared to the state-of-the-art.

1 Introduction

Knowledge Bases (KBs), like Wikipedia, are playing increasingly important roles in numerous entity-based information retrieval tasks. Nevertheless, most KBs are hard to be up-to-date due to their manual maintenances by human editors. As reported in (Frank et al., 2012), there exists a median time lag of 356 days between the day a news article is published and the time that the news is cited in a Wikipedia article dedicated to the entity concerned by the news. The time lag would be reduced if relevant documents could be automatically detected as soon as they are published online and then recommended to the editors. This task is studied as Cumulative Citation Recommendation (CCR). Formally, given a set of KB entities, CCR is to filter relevant documents from a stream corpus and evaluate their citation-worthiness to the target entities.

A variety of supervised approaches (e.g., classification, learning to rank) have been employed and achieved promising results (Wang et al., 2013; Balog and Ramampiaro, 2013; Balog et al., 2013). Nevertheless, most of them leverage all features indiscriminately to build a global relevance model, which leads to unsatisfactory performance. The documents can offer some prior knowledge, which is named as **type** in this paper. The type is the prior knowledge embedded in the document that impacts on the probability of its being recommended to KBs. For instance, when dealing with a document on "music" topic, we would like to have less weights put on a politician entity because this document is not likely to related to it, but more often related to musicians or musical bands. Besides, the source of a document impacts on the recommendation strategies too. A document from news agencies is more reliable and citable than the one from social websites even if they state an identical story about the target KB entity. Hence we consider two kinds of document features to model the prior type knowledge: (1) topic-based features, and (2) source-based features.

This paper proposes a latent document type discriminative mixture model for CCR. We introduce an intermediate latent layer to model latent document types and define a joint distribution over the document-entity pairs and latent document-types on the observation data. The aim is to achieve a discriminative mixture model that is expected to outperform the global relevance model.

To the best of our knowledge, this is the first research work that leverages prior knowledge embedded in documents to improve CCR perfor-

*This work was partially performed when the first author was visiting Purdue University and Microsoft Research Asia.
†Corresponding Author

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 561–566,
Lisbon, Portugal, 17-21 September 2015.

mance. An extensive set of experiments conducted on TREC-KBA-2013 dataset has demonstrated the effectiveness of the proposed mixture model.

2 Discriminative Models for CCR

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 conditional probability of relevance $P(r|e,d)$ with respect to an entity-document pair (e,d). Each (e,d) is represented as a feature vector $\mathbf{f}(e,d) = (f_1(e,d), \cdots, f_K(e,d))$, where K is the dimension of the entity-document feature vector. Moreover, to model the hidden document type, each document is represented as an document-type feature vector $\mathbf{g(d)} = (g_1(d), \cdots, g_L(d))$, where L indicates the dimension of the document-type feature vector.

2.1 Global Model

This paper utilizes logistic regression to estimate the conditional probability $P(r|e,d)$, where $r(r \in \{1,-1\})$ is a binary label indicating the relevance of an entity-document pair (e,d). The value of r is 1 if the document d is related to the entity e, otherwise $r = -1$. Formally, the parametric form of $P(r{=}1|e,d)$ is expressed as $P(r{=}1|e,d) = \delta(\sum_{i=1}^{K} \omega_i f_i(e,d))$, where $\delta(x)$ is the standard logistic function, ω_i is the combination parameter for the ith feature. It is easy to derive 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 $P(r|e,d){=}\delta(r \sum_{i=1}^{K} \omega_i f_i(e,d))$. This model is denoted as **GM** in this paper. Several previous approaches can be deemed as global models adopting different classification functions such as decision trees (Wang et al., 2013) and Support Vector Machine (SVM) (Bonnefoy et al., 2013).

2.2 Latent Document Type Model

In GM, a fixed set of combination weights (i.e., ω) are learned to optimize the overall performance for all entity-document pairs. However, the best combination strategy for a given pair is not always the best for the others since both the documents and entities are heterogeneous. Therefore, we may benefit from developing a document type dependent model in which we choose the combination strategy individually for each document type to optimize the performance for specific document types. Since it is not feasible to determine the proper combination strategy for each document type, we need to classify documents into one of several types. The combination strategy is then tuned to optimize average performance for documents within the same type.

We propose a latent document type model (**LDTM**) by introducing an intermediate layer to capture the underlying type information in documents. A latent variable z is utilized to indicate which type the combination weights ω_z are drawn from. The choice of z is determined by the document d. The joint probability of relevance r and the latent variable z is represented as $P(r,z|e,d;\alpha,\omega){=}P(z|d;\alpha)P(r|e,d,z;\omega)$, where $P(z|d;\alpha)$ is the mixing coefficient, denoting the probability of choosing the hidden type z given document d, and α is the corresponding parameter. $P(r|e,d,z;\omega)$ denotes the discriminative component which takes a logistic function. By marginalizing out z, we obtain

$$P(r|e,d;\alpha,\omega) = \sum_{z}^{N_z} P(z|d;\alpha)\delta\Big(r\sum_{i=1}^{K}\omega_{zi}f_i(e,d)\Big) \quad (1)$$

where ω_{zi} is the weight for the ith entry in the feature vector under the hidden variable z. We adopt a soft-max function $\frac{1}{Z_d}\exp(\sum_{j=1}^{L}\alpha_j^z g_j(d))$ to model $P(z|d;\alpha)$, and Z_d is the normalization factor that scaled the exponential function to be a probability distribution. In this representation, each document d is denoted by a bag of document type features $(g_1(d), \cdots, g_L(d))$. By plugging the soft-max function into Equation (1), we have

$$P(r|e,d;\alpha,\omega){=} \frac{1}{Z_d}\sum_{z=1}^{N_z}\exp\Big(\sum_{j=1}^{L_z}\alpha_{zj}g_j(d)\Big)\delta\Big(r\sum_{i=1}^{K}\omega_{zi}f_i(e,d)\Big) \quad (2)$$

Suppose entity-document pairs in training set are represented as $\mathcal{T}{=}\{(d_u,e_v)\}$, and $\mathcal{R}{=}\{r_{uv}\}$ denotes the corresponding relevance judgment of (d_u,e_v), where $u = 1, \cdots, M$ and $v = 1, \cdots, N$. Assume training instances in $\mathcal{T}$ are independently generated, the conditional likelihood of training data is written as

$$P(\mathcal{R}|\mathcal{T}) = \prod_{u=1}^{M}\prod_{v=1}^{N} P(r_{uv}|e_u,d_v) \quad (3)$$

2.3 Parameter Estimation

The parameters (i.e., ω and α) can be estimated by maximizing the data log-likelihood $\mathcal{L}(\omega, \alpha)$, which is the form of logarithm of Equation (3). A typical parameter estimation method is to use Expectation-Maximization (EM) algorithm by iterating E-step and M-step continuously until convergence. The E-step is derived by computing the posterior probability of z given d_u and e_v, which is denoted as $P(z|d_u, e_v)$.

$$P(z|d_u, e_v) = \frac{\exp\Big(\sum_{j=1}^{L_z}\alpha_{zj}g_j(d_u)\Big)\delta\Big(r_{uv}\sum_{i=1}^{K}\omega_{zi}f_i(d_u, e_v)\Big)}{\sum_z \exp\Big(\sum_{j=1}^{L_z}\alpha_{zj}g_j(d_u)\Big)\delta\Big(r_{uv}\sum_{i=1}^{K}\omega_{zi}f_i(d_u, e_v)\Big)} \tag{4}$$

In M-step, we can obtain the following parameter update rules.

$$\omega_z^* = \arg\max_{\omega_z} \sum_{uv} P(z|d_u, e_v) \log\Big(\delta\Big(\sum_{i=1}^{K}\omega_{zi}f_i(d_u, e_v)\Big)\Big)$$

$$\alpha_z^* = \arg\max_{\alpha_z} \sum_u \Big(\sum_v P(z|d_u, e_v)\Big) \log\Big(\frac{1}{Z_{d_u}}\exp\Big(\sum_{j=1}^{L_z}\alpha_{zj}g_j(d_u)\Big)\Big) \tag{5}$$

To optimize Equation (5), we employ the `minFunc` toolkit[1] using Quasi-Newton strategy. We adopt Akaike Information Criteria (AIC) to determine the number of latent variables (Fang et al., 2010), which is calculated as $2m - 2\mathcal{L}(\omega, \alpha)$, where m is the number of parameters in the model.

LDTM holds two advantages over GM. (1) The combination parameters vary across various document types and hence lead to a gain of flexibility; (2) It offers probabilistic semantics for the latent document types and thus documents can be associated with multiple types.

3 Features

This section presents the two types of features used in the discriminative models. Entity-document features (i.e., $\mathbf{f}(e, d)$) are used in the discriminative components of GM and LDTM. In addition, LDTM requires document-type features (i.e., $\mathbf{g}(e)$) to learn the mixing coefficients in the mixture component.

Since our goal is not to develop new entity-document features, we adopt the identical entity-document feature set proposed in our previous work (Wang et al., 2013; Wang et al., 2015a; Wang et al., 2015b), which has been proved effective.

In terms of document-type features, we consider two kinds of prior knowledge embedded in documents to model the correlations between documents and their latent types.

Topic-based features One prior knowledge to model a document's latent type is its intrinsic topics. As we have claimed, documents with one or more obvious topics are more likely to be recommended to KB than those without any explicit topic. We capture the underlying topics in documents with word co-occurrences. After removing stop words, we represent each document as a feature vector with the bag-of-words model, where word weights are determined by TF-IDF scheme.

Source-based features The source of a document is another prior knowledge to evaluate the probability of the document's being recommended to KBs. We leverage a "bag-of-sources" model to represent each document as source-based feature vector, and term weights are determined by binary occurrence scheme. Please note that the sources are organized hierarchically. For example, *mainstream_news* is a sub-source of *news*.

4 Experiments

4.1 Dataset

We utilize TREC-KBA-2013 dataset[2] as our experimental dataset. The dataset is composed of a temporally stream corpus and a target KB entity set. The stream corpus contains nearly 1 billion documents crawled from 10 sources: **news**, **mainstream_news**, **social**, **weblog**, **linking**, **arxiv**, **classified**, **reviews**, **forum** and **memetracker**[3]. The corpus has 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. The entity set is composed of 121 Wikipedia entities and 20 Twitter entities.

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

[2] `http://trec-kba.org/kba-stream-corpus-2013.shtml`

[3] `http://www.memetracker.org/`

Each entity-document pair is labeled as one of the 4 relevance levels: (i) **Vital**, timely information about the entity's current state, actions, or situation. This would motivate a change to an already up-to-date KB article. (ii) **Useful**, possibly citable but not timely, e.g., background biography, secondary source information. (iii) **Neural**, informative but not citable, e.g., tertiary source like Wikipedia article itself. and (iv) **Garbage**, no information about the target entity could be learned from the document, e.g., spam. Annotation details of the dataset are presented in Table 1.

	Range	Vital	Useful	Neutral	Garbage
Train	2011.10 $\sim$ 2012.02	1696	2121	1030	1702
Test	2012.03 $\sim$ 2013.02	5630	11579	3379	10543

Table 1: Annotation details of TREC-KBA-2013 dataset.

4.2 Evaluation Scenarios

According to different granularity settings, we evaluate the proposed models in two scenarios: (i) **Vital Only**. Only vital entity-document pairs are treated as positive instances. (ii) **Vital + Useful**. Both vital and useful entity-document pairs are treated as positive instances.

4.3 Comparison Methods

We conduct extensive comparisons with the following methods.

- Global Model (**GM**). The global discriminative model introduced in section 2.1.
- Source-based Latent Document Type Model (**src_LDTM**). A variant of LDTM that utilizes source-based features as document-type features.
- Topic-based Latent Document Type Model (**topic_LDTM**). A variant of LDTM that utilizes topic-based features as document-type features.
- Combination Latent Document Type Model (**combine_LDTM**). This approach utilizes source-based and topic-based features together as document-type features. In our experimental setting, we simply union the two feature vectors together into an integral feature vector.

For reference, we also include three top-ranked approaches in TREC-KBA-2013 track.

- **BIT-MSRA** (Wang et al., 2013). A global random forests classification method, the first place approach in TREC-KBA-2013 track.
- **UDEL** (Liu et al., 2013). An entity-centric query expansion approach, the second place approach in TREC-KBA-2013 track.
- **Official Baseline** (Frank et al., 2013). A strong baseline in which human annotators go through target entities and came up with a list of keywords for filtering vital documents.

4.4 Results and Discussion

Improving precision is harder than improving recall for CCR (Frank et al., 2013). Therefore, we care more about recommendation quality of CCR. Precision and overall accuracy are adopted as metrics to evaluate different approaches. All the metrics are computed in the test pool of all entity-document pairs. The results are reported in Table 2. In comparison to the baselines listed in the 2nd block of Table 2, our mixture models achieve higher or competitive precision and accuracy in both scenarios considerably. Compared with the official baseline, our best mixture model improves precision about 28%. In both scenarios, the variants of LDTM outperform GM on precision and accuracy, which validates our motivations that (i) introducing document latent types in mixture model can enhance the recommendation quality, and (ii) source-based and topic-based features can capture the hidden type information of documents.

Methods	Vital Only		Vital + Useful	
	P	Accu	P	Accu
Official Baseline	.171	.175	.540	.532
BIT-MSRA	.214	.445	.589	**.615**
UDEL	.169	.259	.573	.579
GM	.218	.587	.604	.565
src_LDTM	.273	**.763**	.626	.607
topic_LDTM	.293	.755	**.643**	.609
combine_LDTM	**.299**	.751	.633	**.611**

Table 2: Overall results of evaluated methods. Best scores are typeset boldface.

Moreover, topic_LDTM generally performs better than src_LDTM in both scenarios, which meets our expectation because topic-based features have far more dimensions than source-based features. However, even if source-based feature vector holds a few dimensions (10 in our experiments),

src_LDT improves the precision on the basis of GM. Thus, the precision can be enhanced further if we can develop more valuable features to represent the underlying document types. The combination variant of LDTM achieve the best precision in **Vital Only** scenario and the best accuracy in **Vital + Useful** scenario. The naïve combination strategy of two types of features can improve the performance but not stable, so we need find better combination strategies.

For all variant of the LDTM, the number of latent types determined by AIC are reported in Table 3. The optimal number of latent types in **Vital + Useful** is more than that in **Vital Only**. This reveals that the types of **Vital** documents for entities have more restrictions than **Useful** documents, either by topics or by sources. In addition, the optimal number of latent topics is more than that of latent sources, which also follows our intuition that topic-based features holding more dimensions than source-based features. Since we employ a naïve combination strategy for the two types of features, the number of latent types of combine_LDTM is more close to topic_LDTM, which possesses more features than src_LDTM.

Model	Vital	Vital + Useful
src_LDTM	6	7
topic_LDTM	9	15
combine_LDTM	14	15

Table 3: Number of latent types determined by AIC.

5 Related Work

There are three kinds of approaches developed for CCR in previous work: query expansion (Liu et al., 2013; Wang et al., 2013), classification such as SVM (Kjersten and McNamee, 2012) and Random Forest classifier (Bonnefoy et al., 2013; Balog et al., 2013), and learning to rank approaches (Wang et al., 2013; Balog and Ramampiaro, 2013). Transfer learning is utilized to transfer the keyword importance learned from training pairs to query pairs (Zhou and Chang, 2013).

However, some highly supervised methods require training instances for each entity to build a relevance model, limiting their scalabilities. A compromised solution is to build a global discriminative model with all features indifferently. We spotlight document-type features and study the impacts of them in discriminative mixture models. Mixture model has been applied and proved effective in multiple information retrieval tasks, such as expert search (Fang et al., 2010) and federated search (Hong and Si, 2012). By learning flexible combination weights for different types of training instances, mixture model can outperform global models with fixed weights for all instances.

6 Conclusion

Cumulative Citation Recommendation (CCR) is an important task to automatically detect citation-worthy documents from volume text streams for knowledge base entities. We study CCR as a classification problem and propose a latent document type model (LDTM) through introducing a latent layer in a discriminative model to capture the correlations between documents and their intrinsic types. Two variants of LDTM are implemented by modeling the latent types with document source-based and topic-based features respectively. Experimental results on TREC-KBA-2013 dataset demonstrate that our mixture model can improve CCR performance significantly, especially on precision and accuracy, revealing the advantage of LDTM in enhancing recommendation quality of citation-worthy documents.

For future work, we wish to explore more useful document-type features and apply more proper combination strategies to improve the latent document type model.

Acknowledgement

The authors would like to thank Fei Sun, Qifan Wang and Chen Shao for their valuable suggestions and the anonymous reviewers for their helpful comments. This work is funded by the National Program on Key Basic Research Project (973 Program, Grant No. 2013CB329600), National Natural Science Foundation of China (NSFC, Grant Nos. 61472040 and 60873237), and Beijing Higher Education Young Elite Teacher Project (Grant No. YETP1198).

References

Krisztian Balog and Heri Ramampiaro. 2013. Cumulative citation recommendation: classification vs. ranking. In *SIGIR*, pages 941–944. ACM.

Krisztian Balog, Heri Ramampiaro, Naimdjon Takhirov, and Kjetil Nørvåg. 2013. Multi-step classification approaches to cumulative citation recommendation. In *OAIR*, pages 121–128.

Ludovic Bonnefoy, Vincent Bouvier, and Patrice Bellot. 2013. A weakly-supervised detection of entity central documents in a stream. In *SIGIR*, pages 769–772.

Yi Fang, Luo Si, and Aditya P. Mathur. 2010. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *SIGIR*, pages 683–690. ACM.

J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Re, and I. Soboroff. 2012. Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In *TREC*.

John Frank, Steve J. Bauer, Max Kleiman-Weiner, Daniel A. Roberts, Nilesh Triouraneni, Ce Zhang, and Christopher Rè. 2013. Evaluating stream filtering for entity profile updates for trec 2013. In *TREC*.

Dzung Hong and Luo Si. 2012. Mixture model with multiple centralized retrieval algorithms for result merging in federated search. In *SIGIR*, pages 821–830. ACM.

Brain Kjersten and Paul McNamee. 2012. The hltcoe approach to the trec 2012 kba track. In *TREC*.

Xitong Liu, Jeffrey Darko, and Hui Fang. 2013. A related entity based approach for knowledge base acceleration. In *TREC*.

Jingang Wang, Dandan Song, Chin-Yew Lin, and Lejian Liao. 2013. Bit and msra at trec kba ccr track 2013. *TREC*.

Jingang Wang, Lejian Liao, Dandan Song, Lerong Ma, Chin-Yew Lin, and Yong Rui. 2015a. Resorting relevance evidences to cumulative citation recommendation for knowledge base acceleration. In *Web-Age Information Management*, volume 9098 of *Lecture Notes in Computer Science*, pages 169–180. Springer International Publishing.

Jingang Wang, Dandan Song, Qifan Wang, Zhiwei Zhang, Luo Si, Lejian Liao, and Chin-Yew Lin. 2015b. An entity class-depedent discriminative mixture model for cumulative citation recommendation. In *SIGIR*. ACM.

Mianwei Zhou and Kevin Chen-Chuan Chang. 2013. Entity-centric document filtering: boosting feature mapping through meta-features. In *CIKM*, pages 119–128. ACM.