NERank: Ranking Named Entities in Document Collections

Chengyu Wang, Rong Zhang, Xiaofeng He*, Aoying Zhou
Institute for Data Science and Engineering
East China Normal University, Shanghai, China
chywang2013@gmail.com, {rzhang,xfhe,ayzhou}@sei.ecnu.edu.cn

ABSTRACT

While most of the entity ranking research focuses on Web corpora with user queries as input, little has been done to rank entities directly from documents. We propose a ranking algorithm **NERank** to address this issue. **NERank** employs a random walk process on a weighted tripartite graph mined from the document collection. We evaluate **NERank** over real-life document datasets and compare it with baselines. Experimental results show the effectiveness of our method.

Keywords

entity ranking; random walk; tripartite graph

1. INTRODUCTION

Ranking problems have been extensively studied to bring order to varying types of objects, such as Web pages, products and textual units. With the number of entities increasing rapidly on the Web, the problem of Entity Ranking (ER) has drawn much attention. For example, ER tracks have been conducted in INEX and TREC since 2006 and 2009, respectively, to rank entities from semi-structured Web corpora given a query topic.

In traditional ER tasks, the rank of entities is measured by the relevance between a query topic and entities with contextual information. In this paper, we consider a different problem: *ranking entities in document collections based on the importance of entities in documents*. The challenge is that the ranking order of entities should be determined by the contents of the document collection, with no additional data sources available.

The task of ER introduced in this paper is interesting for several reasons: (1) it automatically identifies important entities in plain text; (2) it facilitates entity recommendation in Web search; and (3) it potentially improves the performance of other tasks such as knowledge base population by extracting and ranking entities from the Web.

In this paper, we propose a graph-based ranking algorithm **NERank** to solve this task. Given a document collection as input, we mine latent topics and model the semantic relatedness between documents, topics and entities in a weighted tripartite graph. We design a ranking function to estimate prior ranks of topics, and propagate the ranks along paths in the tripartite graph via a modified random walk process. Details of **NERank** are presented in Section 2. Experiments are shown in Section 3.

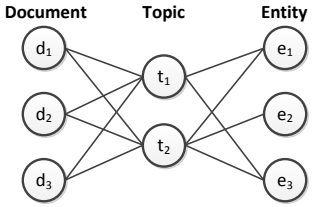


Figure 1: A simple tripartite graph for NERank.

2. PROPOSED APPROACH

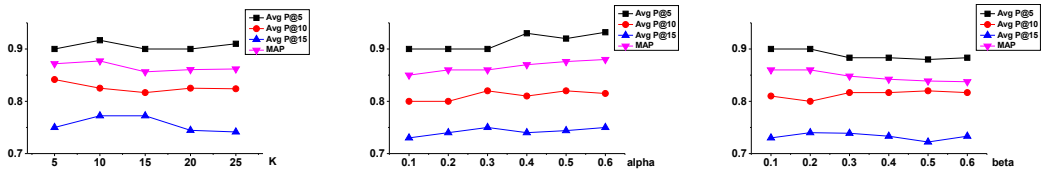
We take a collection of documents (denoted as D) as input. Let M denote the collection of entity mentions detected from D via NER. E denotes the collection of normalized named entities in D. An entity normalization procedure maps each mention $m \in M$ to its normalized form $e \in E$. ER assigns each entity $e \in E$ a rank $r(e)$ to represent the relative importance in D. **NERank** addresses the task of ER using three modules, discussed as follows:

Data Preprocessing and Graph Construction. We perform NER and named entity normalization to generate the entity set M, and map each $m \in M$ to the normalized form $e \in E$. Implementation details are described in [1]. We employ LDA to mine the latent topics T in D. Rather than treating a document as a single word collection, we model a document as the union of common words (words that do not refer to any named entities) and normalized named entities. Document-topic distribution matrix $\mathbf{\Theta}$ and topic-word distribution matrix $\mathbf{\Phi}$ are estimated in LDA.

We employ a weighted, tripartite graph to model the semantic relationships among documents, topics and entities (illustrated in Fig. 1). In the graph, there are three types of nodes (i.e., documents D, topics T and entities E), and two types of undirected edges (i.e., document-topic and topic-entity edges). We use weights of the edges to represent how close the semantic relatedness is between corresponding nodes. For a document-topic edge (d_i, t_j), we define the weight $w(d_i, t_j) = \theta_{i,j}$ where $\theta_{i,j}$ is the element in the i^{th}

*Corresponding author.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889348.



(a) Varying K ($\alpha = \beta = 0.4$) (b) Varying α ($K = 10, \beta = 0.2$) (c) Varying β ($K = 10, \alpha = 0.2$)

Figure 2: Evaluation results under different parameter settings.

row and the j^{th} column of $\mathbf{\Theta}$. We remove columns for distributions of all the common words in $\mathbf{\Phi}$, and denote the rest part of the matrix as $\hat{\mathbf{\Phi}}$. For a topic-entity edge (t_i, e_j), we have $w(t_i, e_j) = \hat{\phi}_{i,j}$ where $\hat{\phi}_{i,j}$ is the element in the i^{th} row and the j^{th} column of $\hat{\mathbf{\Phi}}$.

Prior Topic Rank Estimation. With the help of LDA, we estimate the ranks of topics by studying the distributions of topics. Let $r_0(t)$ denote the prior rank for topic t. We propose three quality metrics for each topic $t \in T$. Quality metrics include *prior probability* (the probability that topic t is discussed in D) $pr(t)$, *entity richness* (the proportion of named entities in words related to topic t) $er(t)$ and *topic specificity* (whether the topic is specific about certain aspects or only provides background information) $ts(t)$. We combine the three quality metrics in a linear function to compute the prior ranks of topics, defined as:

$$r_0(t) = \frac{1}{Z}(w_1 \cdot pr(t) + w_2 \cdot er(t) + w_3 \cdot ts(t))$$

where $Z = \sum_{t' \in T} r_0(t')$ is a normalization factor. $\forall i, w_i > 0$ and $w_1 + w_2 + w_3 = 1$. The weights can be learned using the max-margin technique introduced in [3].

Random Walk Process. We design a random walk-based algorithm according to the link structure of the tripartite graph. The random surfer begins by selecting a topic node $t_i \in T$ with probability $r_0(t_i)$ as the starting point. We define α and β as tuning parameters where $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$. Denote $x \to y$ as the walk from x to y. The random surfer makes one of the following three transfers:

1. With prob. α, the random surfer walks through the path $t_i \to d_j \to t_k$. $d_j \in D$ is selected with prob. $\frac{\theta_{j,i}}{\sum_{d_k \in D} \theta_{k,i}}$. Next, $t_k \in T$ is selected with prob. $\theta_{j,k}$.

2. With prob. β, the random surfer walks through the path $t_i \to e_j \to t_k$. $e_j \in E$ is selected with prob. $\frac{\hat{\phi}_{i,j}}{\sum_{e_k \in E} \hat{\phi}_{i,k}}$. Next, $t_k \in T$ is selected with prob. $\frac{\hat{\phi}_{k,j}}{\sum_{e_m \in E} \hat{\phi}_{k,m}}$.

3. With prob. $1 - \alpha - \beta$, the random surfer jumps to a topic node t_j. t_j is selected with prob. $r_0(t_j)$.

This process can be repeated iteratively until the system reaches equilibrium. Each entity e_i will receive a score $s(e_i)$, indicating the number of visits by random surfers. The rank of entity e_i is computed as $r(e_i) = \frac{s(e_i)}{\sum_{e_j \in E} s(e_j)}$.

3. EXPERIMENTS

Datasets. We use two publicly available datasets: **TimelineData** [5] and **CrisisData** [4]. They consist of document collections related to major international events, such as Egypt revolution, Syria war, etc. We conduct separate experiment on all the document collections in these datasets.

Evaluation. For each document collection, 15 entities are annotated as key entities by humans that should be extracted by the ranking algorithm. We employ Precision@K ($K = 5, 10, 15$) and Average Precision as the evaluation metrics. For all the document collections, we report the average values of these metrics (i.e., Average Precision@K and MAP). We also test whether our method is better than other methods and report the significance level by p-value.

Table 1: Evaluation results for different methods. ($\star$: p-value$\leq$0.05)

Method	AvgP@5	AvgP@10	AvgP@15	MAP
TF-IDF	0.85*	0.79*	0.73*	0.81*
TextRank	0.87*	0.83	0.73*	0.83*
NERank$_{Uni}$	0.80*	0.75*	0.71*	0.78*
NERank$_{\alpha=0}$	0.72*	0.61*	0.51*	0.62*
NERank$_{Full}$	**0.92**	**0.87**	**0.79**	**0.89**

Parameter Tuning. We tune parameters in **NERank**, namely, number of topics in LDA (K) and parameters for random walk (α and β). In Fig. 2, we present the experimental results when we vary only one parameter at each time. It can be seen that **NERank** is not sensitive to the changes of parameters.

Comparison. We compare our method against baselines. The results are shown in Table 1. We compute the score for each entity using two baselines **TF-IDF** and **TextRank** [2]. We also evaluate two variant of the purposed approach: **NERank**$_{Uni}$ (which assigns prior topic ranks uniformly) and **NERank**$_{\alpha=0}$ (which sets $\alpha = 0$ in random walk and thus ignores the semantic relatedness between documents and topics). The results show that the purposed approach **NERank**$_{Full}$ outperforms all the other baselines.

Acknowledgments

This work is partially supported by NSFC under Grant No. 61402180, and the Natural Science Foundation of Shanghai under Grant No. 14ZR1412600.

4. REFERENCES

[1] V. Jijkoun, M. A. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. In *AND*, pages 23–30, 2008.

[2] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *EMNLP*, pages 404–411, 2004.

[3] W. Shen, J. Wang, P. Luo, and M. Wang. LINDEN: linking named entities with knowledge base via semantic knowledge. In *WWW*, pages 449–458, 2012.

[4] G. B. Tran, M. Alrifai, and E. Herder. Timeline summarization from relevant headlines. In *ECIR*, pages 245–256, 2015.

[5] G. B. Tran, M. Alrifai, and D. Q. Nguyen. Predicting relevant news events for timeline summaries. In *WWW*, pages 91–92, 2013.