Short-Text Conceptualization Based on A Co-Ranking Framework via Lexical Knowledge Base

Yashen Wang[1]*

[1]China Academy of Electronics and Information Technology of CETC, Beijing, China
yashen_wang@126.com

Abstract. The problem of short-text conceptualization is important, and has attracted increasing attention. Recent probabilistic algorithms have demonstrated remarkable successes. However most of them are limited to the assumption that all the observed terms in given short-text are conditionally independent, ignoring the interaction among terms (and concepts), as well as the beneficial reactions from concepts to terms. To overcome these problems, recently some co-rank paradigms are proposed, unfortunately neither they fails to integrate the co-occurrence feature nor they fails to utilize the semantic similarity implicit in the lexical knowledge base. Therefore, previous works could not release robust concept representation. Faced with this problem, this paper proposes a novel framework based on both statistic information (e.g., co-occurrence feature in large-scale corpus) and semantic information (e.g., semantic similarity in lexical knowledge base), for co-ranking terms and their corresponding concepts simultaneously, This co-ranking framework utilizes several graphs: the concept graph, the term graph and the subordination graph. The experimental results show that our method achieves higher accuracy and efficiency in short-text conceptualization than the state-of-the-art algorithms.

Keywords: Conceptualization · Co-Ranking · Lexical Knowledge Base

1 Introduction

Shot-text conceptualization, is an interesting task to infer the most likely concepts for terms in the short-text, which could help better make sense of text data, and extend the texts with categorical or topical information [1,30,9,28,23,27]. It is a task to map a piece of short-text to a set of open domain concepts with different granularities[10,12,26,8]. Recent works on short-text understanding have put more emphasis on using signals from lexical knowledge bases to assist short-text conceptualization [9,32,28], and achieve great success. Many probabilistic (graph-based) algorithms have been proposed [21,30,23,10]. Generally, these kind of algorithms is closely integrated with the knowledge base, and the knowledge base has been demonstrated to be used to helping short-text understanding [29,32,28]. Given a short-text as input, we map each term to the corresponding candidate concepts defined in lexical knowledge base (e.g., Probase), and therefore a semantic graph is constructed based on the terms, concepts and the links among them. Note that, This semantic graph is heterogeneous [5,20,19], including three

* The corresponding author.

sub-graph: (i) the concept graph G_C connecting concepts (defined in the lexical knowledge base); (ii) the term graph G_T connecting terms (embedded in the short-text), and (iii) the subordination graph G_{TC} that ties the two previous graphs together.

As concluded in [10], after mapping the terms $T = \{t_i | i = 1, \ldots, n_T\}$ in given short-text to some candidate concepts in lexical knowledge base (e.g., Probase), previous works aim at estimating optimal set of concepts $C = \{c_j | j = 1, \ldots, k_C\}$, which maximizes conditional probabilities $P(c_j|T) \propto P(c_j) \prod_{i=1}^{n_T} P(t_i|c_j)$[1] based on the co-occurrence of terms and concepts under Naive Bayes assumption [21,30]. In fact, they fail in mining the holistic concept-set for the entire short-text. The reasons are discussed as follows: (i) They assume that all the observed terms (and concepts) are conditionally independent, ignoring the beneficial reactions from concept to terms, which could reflect the global concepts, and simply regard the multiplication of conditional probabilities from each term as the likelihood of concept c_j [23,21]. (ii) Recently some co-rank paradigms are proposed to investigate the beneficial reactions among terms and concepts, unfortunately neither they fails to integrate the co-occurrence feature [17] nor they fails to utilize the semantic similarity implicit in the lexical knowledge base [10].

So as to overcome these problems, we must: (i) devise a framework that enables the signals (i.e., terms and concepts) to fully interplay to derive solid conceptualization for short-text; and (ii) combines *global* statistic information (e.g., co-occurrence feature from large corpus), *local* information (heuristic information implicit in context, i.e., correlation function) and *manual-defined knowledge* (semantic similarity in lexical knowledge base). Therefore, we propose a framework to co-rank terms and their concepts simultaneously in the concept graph (G_C), the term graph (G_T) and the subordination graph (G_{TC}). As a result, improved rankings of terms and their concepts depend on each other in a mutually reinforcing way, thus taking advantage of the additional information implicit in such heterogeneous graph of terms and concepts. The main intuition behind the co-ranking strategy is that, there is a mutually reinforcing relationship among concepts and terms that could be reflected in the rankings.

2 Preliminary

2.1 Problem Definition

Following [26], we define the notation "concept" as a set or class of "entities" or "things" within a domain, such that words belonging to similar classes get similar representations. Probase [32] is used in our study as lexical knowledge base. Probase is widely used in research about short-text understanding [31,22,24] and text representation [10,28]. Probase uses an automatic and iterative procedure to extract concept knowledge from 1.68 billion Web pages. It contains 2.36 millions of open domain terms. Each term is a concept, an instance, or both. Meanwhile, it provides around 14 millions relationships with two kinds of important knowledge related to concepts: concept-attribute co-occurrence (isAttrbuteOf) and concept-instance co-occurrence

[1] Notation n_T indicates the number of the terms occurring in this heterogeneous semantic graph, which will be discussed later

(isA). Moreover, Probase provides huge number of high-quality and robust concepts without builds.

Given a short-text $S = \{t_i | i = 1, \ldots, n_T\}$, wherein t_i denotes a term, we could obtain the following results via short-text conceptualization: (i) concept distribution $\phi_C = \{\langle c_i, RS_C(i)\rangle | i = 1, \cdots, k_C\}$ from lexical knowledge base , wherein $RS_C(i)$ indicates the ranking score of concept c_i representing the importance of concept c_i contributing to model the entire semantic of the given short-text S (details in Section 3); and (ii) key-term distribution $\phi_T = \{\langle t_j, RS_T(j)\rangle | j = 1, \cdots, k_T\}$, wherein $RS_T(j)$ indicates the ranking score of term t_j representing the importance of term t_j contributing to model the entire semantic of the given short-text S. Through the above-mentioned definition, the essence of short text conceptualization is to map a given short-text to a concept space. This mapping process could filter out the incorrect concepts that are not suitable for the current given context, and then achieve the semantic disambiguation of polysemy.

2.2 Heterogeneous Semantic Graph

As discussed in the begining section, the proposed co-ranking framework operates on a heterogeneous semantic graph, which consists of three sub-graphs. Overall, we denote the heterogeneous semantic graph as $G = (V_C \cup V_T, E_{CC} \cup E_{TC} \cup E_{TT})$. Wherein, V_C is the set of candidate concepts with size of $n_C = |V_C|$, and V_T is the set of terms with size of $n_T = |V_T|$. E_{CC} is the set of links representing correlation ties among concepts, E_{TT} is the set of links among terms established by their co-occurrence relations, and E_{TC} is the set of links representing the subordination relations among terms and concepts. The overall heterogeneous semantic graph G is composed of three sub-graphs: (i) the Concept Graph $G_C = (V_C, E_C)$ respecting concepts, (ii) the term-correlation graph $G_T = (V_T, E_T)$ respecting terms, and (iii) the bipartite subordination graph $G_{TC} = (V_{TC}, E_{TC})$ that ties concepts (in G_C) and terms (in G_T) together.

2.3 Affinity Matrix

Overall, the proposed co-ranking framework is controlled by four affinity matrices. Note that, the affinity matrix is also reviewed as the transition matrix in Markov chain and is a stochastic matrix prescribing the transition probabilities from one vertex (concept or term) to the next, as discussed in the following Section 3. The affinity matrix $\mathbf{M}$ of G is defined as follows:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M_{CC}} & \mathbf{M_{CT}} \\ \mathbf{M_{TC}} & \mathbf{M_{TT}} \end{bmatrix} \tag{1}$$

wherein $\mathbf{M}_{CC}$ represents the correlation relationship among concepts (in G_C, defined in Section 2.2), and $\mathbf{M}_{TT}$ represents the co-occurrence relationship among terms (in G_W, defined in Section 2.2). $\mathbf{M}_{CT}$ and $\mathbf{M}_{TC}$ denote bipartite subordination (in G_{TC} , defined in Section 2.2), measuring how likely the given term is assigned with some concepts and vice versa.

Concept Graph (G_C): The Concept Graph $G_C = (V_C, E_C)$, representing the relatedness among candidate concepts associated with the given short-text, is a weighted

undirected graph. The individual concept is denoted as $\{c_i|c_i \in V_C, i = 1, 2, \ldots, n_C\}$. $M_{CC}[i][j]$, the element of $\mathbf{M_{CC}}$, is derived by aggregating the co-occurrences between all instances of the two concepts c_i and c_j and the semantic similarity between the two concepts c_i and c_j. To achieve this goal, we firstly define the semantic similarity between concept c_i and concept c_j, as follows:

$$\mathrm{sim}(c_i, c_j) = \frac{|T_{c_i} \cap T_{c_j}|}{|T_{c_i}|} \tag{2}$$

Wherein, T_{c_i} indicate the set of terms belong to concept c_i defined in lexical knowledge base Probase, and T_{c_j} could be defined in the same way. With efforts above, we could utilize the following equation to define $M_{CC}[i][j]$:

$$M_{CC}[i][j] = \frac{\eta_{CC} \cdot \sum_{t_p \in c_i, t_q \in c_j} n(t_p, t_q) + (1 - \eta_{CC}) \cdot \mathrm{sim}(c_i, c_j)}{\sum_{l=1}^{n_C} [\eta_{CC} \cdot \sum_{t_p \in c_i, t_q \in c_l} n(t_p, t_q) + (1 - \eta_{CC}) \cdot \mathrm{sim}(c_i, c_l)]} \tag{3}$$

Wherein, η_{CC} is the parameter controlling the weights about the co-occurrence feature and the semantic similarity feature. t_p and t_q are terms in vocabulary, and $n(t_p, t_q)$ represents the co-occurrence frequency between them through statistics. Furthermore, we could define correlation function for each pair of c_i and c_j resulting from their co-participation of term t_k:

$$\tau(c_i, c_j, t_k) = \frac{\mathrm{II}(M_{TC}[k][i] \neq 0, M_{TC}[k][j] \neq 0)}{|t_k|(|t_k| - 1)/2} \tag{4}$$

Wherein $\mathrm{II}(M_{TC}[k][i] \neq 0, M_{TC}[k][j] \neq 0)$ is the indicator function of whether term t_k could be mapped to concepts c_i and c_j simultaneously, and $|t_k|$ denotes the number of all the concepts related to t_k. Hence, adding up correlation function from all terms, we obtain:

$$M_{CC}[i][j] = \frac{[\eta_{CC} \cdot \sum_{t_p \in c_i, t_q \in c_j} n(t_p, t_q) + (1 - \eta_{CC}) \cdot \mathrm{sim}(c_i, c_j)] \cdot \sum_{k=1}^{n_T} \tau(c_i, c_j, t_k)}{\sum_{l=1}^{n_C} [\eta_{CC} \cdot \sum_{t_p \in c_i, t_q \in c_l} n(t_p, t_q) + (1 - \eta_{CC}) \cdot \mathrm{sim}(c_i, c_l)] \cdot \sum_{k=1}^{n_T} \tau(c_i, c_l, t_k)} \tag{5}$$

Term Graph (G_T): We segment the given short-text into a set of terms $\{t_i|t_i \in V_T, i = 1, 2, \ldots, n_T\}$, by utilizing the Probase [30] as our lexicon. The Term Graph $G_T = (V_T, E_T)$ is an weighted undirected graph representing co-occurrence relations among terms in given short-text. Similarly, $M_{TT}[i][j]$, the element of $\mathbf{M_{TT}}$, is derived by aggregating the co-occurrences between the two term t_i and t_j and the semantic similarity between them. For given term t_i, we denote its concept set as C_{t_i}, consisting the corresponding concepts deriving from Probase by leveraging single instance conceptualization algorithm [31,28,10]. Therefore, we define the semantic similarity between term t_i and concept t_j, as follows:

$$\mathrm{sim}(t_i, t_j) = \frac{|C_{t_i} \cap C_{t_j}|}{|C_{t_i}|} \tag{6}$$

With efforts above, $M_{TT}[i][j]$ could be defined as follows:

$$M_{TT}[i][j] = \frac{\eta_{TT} \cdot n(t_i, t_j) + (1 - \eta_{TT}) \cdot \mathrm{sim}(t_i, t_j)}{\sum_{k=1}^{n_T} [\eta_{TT} \cdot n(t_i, t_k) + (1 - \eta_{TT}) \cdot \mathrm{sim}(t_i, t_k)]} \tag{7}$$

Wherein, η_{TT} is the parameter controlling the weights about the co-occurrence feature and the semantic similarity feature. Moreover, we could also take local information implicit in this context into consideration. Therefore, we also introduce a correlation function for each pair of t_i and t_j (given concept c_k) to differentiate different attention:

$$\sigma(t_i, t_j, c_k) = \frac{\amalg(M_{TC}[i][k] \neq 0, M_{TC}[j][k] \neq 0)}{|c_k|(|c_k| - 1)/2} \tag{8}$$

Wherein $|c_k|$ denotes the number of all the terms related to concept c_k. Adding up correlation function from all concepts, we obtain

$$M_{TT}[i][j] = \frac{[\eta_{TT} \cdot n(t_i, t_j) + (1 - \eta_{TT}) \cdot \mathrm{sim}(t_i, t_j)] \cdot \sum_{k=1}^{n_C} \sigma(t_i, t_j, c_k)}{\sum_{l=1}^{n_T} [\eta_{TT} \cdot n(t_i, t_l) + (1 - \eta_{TT}) \cdot \mathrm{sim}(t_i, t_l)] \cdot \sum_{k=1}^{n_C} \sigma(t_i, t_l, c_k)} \tag{9}$$

Subordination Graph (G_{TC}): $G_{TC} = (V_{TC}, E_{TC})$ is a weighted bipartite graph representing relationship among terms all of their corresponding concepts and leveraging the previous graphs, wherein $V_{TC} = V_T \cup V_C$. $M_{TC}[i][j]$ represents the link from a term t_i to a concept c_j. We formulate the subordinate degree of term t_i and concept c_j:

$$\mathrm{sub}(t_i, c_j) = \frac{n_{ins}(t_i, c_j)}{\sum_{k=1}^{n_C} n_{ins}(t_i, c_k)} \tag{10}$$

Wherein, $n_{ins}(t_i, c_j)$ is the frequency that term t_i is an instance of concept c_j. Moreover, we also takes local information embedded in the current context into consideration, by introducing an correlation function for pair of term t_i and term t_j (given concept c_k):

$$\varphi(t_i, t_j, c_k) = \frac{\amalg(M_{TC}[i][k] \neq 0, M_{TC}[j][k] \neq 0)}{|t_j|(|t_j| - 1)/2} \tag{11}$$

Furthermore, add up this correlation function from all terms:

$$M_{TC}[i][j] = \frac{\mathrm{sub}(t_i, c_j) * \sum_{k=1}^{n_T} \varphi(t_i, t_k, c_j)}{\sum_{l=1}^{n_C} \mathrm{sub}(t_i, c_l) * \sum_{k=1}^{n_T} \varphi(t_i, t_k, c_l)} \tag{12}$$

As demonstrated in [10], we also assign the normalization of subordinate degree to $M_{CT}[i][j]$ straightly, rather than introducing the correlation function above:

$$M_{CT}[i][j] = \frac{\mathrm{sub}(t_j, c_i)}{\sum_{l=1}^{n_T} \mathrm{sub}(t_l, c_i)} \tag{13}$$

3 The Proposed Co-Ranking Framework

Based on the construction of the heterogeneous semantic graph (Section 2.2) and the affinity matrixes (Section 2.3), the proposed co-ranking framework operates the following iteration procedure, consisting for step in each iteration, on G_C, G_T and G_{TC} mutually to mine the most expressive concepts similar to [10], and when this iteration procedure converges, we choose the top-k_C concepts and top-k_T terms according to descending ranking-scores as final results. The algorithm typically converges when difference between the ranking-scores computed at two successive iterations falls below a presupposed threshold. As shown in Fig. 1, to simultaneously tune intra-class rankings (among homogenesis elements) and inter-class rankings (among heterogeneous elements), a set of asymmetric parameters $\gamma_{CC}, \gamma_{CT}, \gamma_{TC}, \gamma_{TT} \in [0, 1]$ is used to determining the weights of different random walk procedure in different sub-graphs, with the following constraints: $\gamma_{CC} + \gamma_{CT} + \gamma_{TC} + \gamma_{TT} = 1$ (as demonstrated in [10]).

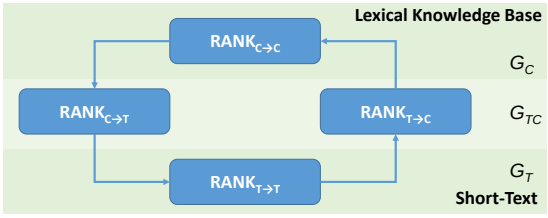

Fig. 1: The proposed co-ranking framework for short-text conceptualization.

Term ranks Concept (RANK$_{T\to C}$ in Fig. 1). The ranking-scores of terms are used to reinforcing the scores of concepts, and these values are initially set as TF-IDF in corpus.

$$RS_C^{(z+1)}(i) = \gamma_{TC} \Sigma_{k=1}^{n_T} M_{TC}[k][i] * RS_T^{(z)}(k) \tag{14}$$

$$\mathbf{RS_C^{(z+1)}} = \mathbf{RS_C^{(z+1)}} / \|\mathbf{RS_C^{(z+1)}}\| \tag{15}$$

Wherein, $\mathbf{RS_C^{(z+1)}}$ and $\mathbf{RS_T^{(z+1)}}$ denote ranking-score vector for concepts and terms in *z+1*-th iteration, and $RS_C^{(z+1)}(i)$ and $RS_T^{(z)}(k)$ denote the ranking-scores of concept c_i and term t_k. To guarantee convergence, $\mathbf{RS_C^{(z)}}$ and $\mathbf{RS_T^{(z)}}$ are normalized after each iteration [7,30,34].

Concept ranks Concept (RANK$_{C\to C}$ in Fig. 1). The ranking-scores of concepts are used to reinforcing the scores of other concepts based on their relevance. Note that, We rank the concept graph G_C following the PageRank paradigm [3], which is somewhat similar to weighted PageRank [6]. Consider a random walk on G_C, and the affinity matrix $\mathbf{M_{CC}}$ could be reviewed as the transition matrix. Note taht, a random walk on a graph is a Markov chain [15], its states being the vertices of the graph

$$RS_C^{(z+1)}(i) = \gamma_{CC}(1-\beta_{CC}+\beta_{CC}*\Sigma_{k=1}^{n_C} M_{CC}[k][i]*RS_C^{(z)}(k)) \quad (16)$$

$$\mathbf{RS_C^{(z+1)}} = \mathbf{RS_C^{(z+1)}}/\|\mathbf{RS_C^{(z+1)}}\| \quad (17)$$

Wherein β_{CC} is the damping factor as used in PageRank, and at each time step with probability $(1-\beta_{CC})$ we stick to random walking and with probability β_{CC} we do not make a usual random walk step, but instead jump to any vertex, chosen uniformly at random.

Concept ranks Term (RANK$_{C\to T}$ in Fig. 1). The ranking-scores of concepts are used to reinforcing the scores of terms.

$$RS_T^{(z+1)}(j) = \gamma_{CT}\Sigma_{k=1}^{n_C} M_{CT}[k][j] * RS_C^{(z)}(k) \quad (18)$$

$$\mathbf{RS_T^{(z+1)}} = \mathbf{RS_T^{(z+1)}}/\|\mathbf{RS_T^{(z+1)}}\| \quad (19)$$

Term ranks Term (RANK$_{T\to T}$ in Fig. 1). The ranking-scores of terms are used to reinforcing the scores of other terms based on their relevance.

$$RS_T^{(z+1)}(j) = \gamma_{TT}(1-\beta+\beta_{TT}*\Sigma_{k=1}^{n_T} M_{TT}[k][j]*RS_T^{(z)}(k)) \quad (20)$$

$$\mathbf{RS_T^{(z+1)}} = \mathbf{RS_T^{(z+1)}}/\|\mathbf{RS_T^{(z+1)}}\| \quad (21)$$

Wherein β_{TT} is also the damping factor as used in PageRank. The fact that there exists a unique solution to Eq. (21) follows from the random walk $\mathbf{M_{TT}}$ being ergodic[2].

4 Experiments and Results

Since there exists no concept-annotated corpus for short-texts, to validate the performance of our co-ranking framework and other state-of-the-art algorithms, we conduct experiments on text clustering task, which is widely used for evaluating text conceptualization [23,21], to evaluate the results.

4.1 Datasets

Following [10], we preprocess the Wikipedia articles to construct corpus **Wiki** for construction of the affinity matrix **M** of the heterogeneous semantic graph G, which contains 3.74 million Wikipedia articles. For text clustering task, we use three datasets: **NewsTitle**, **Twitter**, **WikiFirst** and **TREC**, as follows:

NewsTitle: We extract news titles from a news corpus containing 3.62 million articles searched from Reuters and New York Time. The news articles are classified into six categories: *company*, *religion*, *science*, *traffic*, *politician*, and *sport*. We randomly select 5,000 news titles in each category. The average word count of titles is 9.37.

[2] $\beta_{TT} > 0$ guarantees irreducibility [17], because we can jump to any vertex.

Twitter: We utilize the official tweet collections used in TREC Microblog Task 2013/2014 to construct this dataset. By manually labeling, the dataset contains 41,536 tweets which are in four categories: *food*, *sport*, *entertainment*, and *device/IT company*. We remove the URLs and stop-words. The average length of the tweets is 12.95 words. Because of noise and sparsity, this dataset is more challenging.

WikiFirst: this dataset includes 330,000 Wikipedia articles, which are divided into 110 categories based on the mapping relationship between Wikipedia articles and Freebase topics. For example, Wikipedia articles titled The "Big Bang Theory" are categorized into *Tv_program* in Freebase. Each category contains 3,000 Wikipedia articles. We extract the first sentence of each Wikipedia article to construct this dataset, and the average length of the first sentence is 12.67 words. Note that, this dataset is a challenging data set because of its large number of categories, strong diversity of categories and strong correlation among many categories.

TREC: It is the corpus for question clustering on TREC [14], which is widely used as benchmark. The entire dataset of 5,952 sentences are classified into the six categories: *person*, *entity*, *abbreviation*, *description*, *location* and *numeric*.

4.2 Alternative Algorithms and Experiment Settings

We compare the proposed framework with the following short-text conceptualization algorithms:

BOW: It represents short-text as bag-of-words with the TF-IDF scores [18].

LDA: It represents short-text as its inferred topic distribution [2], and the dimensions of the short-text vector of is number of topics as we presuppose.

IJCAI$_{11}$: [21] proposed a probabilistic framework, which performed a simple co-clustering of concepts and terms by identifying the disjoint cliques, and then derived the most likely concepts using Bayesian inference.

IJCAI$_{11}$+CL: By introducing the clustering strategy, [21] extends **IJCAI$_{11}$**. [21] first mines dense k-exclusive clusters that maximize conditional probability $P(t_i|c_j)$, where the words in the same cluster are considered to belong to the same semantic cluster. Then the algorithm **IJCAI$_{11}$** is implemented on each semantic cluster to complete short-text conceptualization.

IJCAI$_{15}$: Taking verbs and adjectives into consideration, [30] conceptualized terms using a random-walk based iterative algorithm.

RW: It is a pure random walk variant of **IJCAI$_{15}$**, without adjusting the weights on links during whole procedure.

Co-Rank$_{IP}$: [10] ranks the concepts and terms simultaneously in an iterative procedure based on a co-ranking framework.

Co-Rank (Ours): By leveraging concept-based similarity among terms, the proposed co-ranking framework boosts [10], and achieves the goal of combining global statistic information (e.g., co-occurrence feature from large corpus), local information (heuristic information implicit in context, i.e., correlation function) and manual-defined knowledge (semantic similarity in lexical knowledge base).

Co-Rank$_{AD}$: A co-ranking framework by simply coupling two random walks [34], which separately rank different type of vertices under PageRank [3].

With the limitation of space, we briefly describe the experimental settings here. The dimensions of vector in **BOW** is 25,000, which releases the optimal experimental results. For **Co-Rank**, **Co-Rank$_{\mathbf{AD}}$**, **Co-Rank$_{\mathbf{HITS}}$**, **IJCAI$_{\mathbf{11}}$**, **IJCAI$_{\mathbf{11}}$+CL**, **IJCAI$_{\mathbf{15}}$** and **RW**, we select 5,000 concepts as features in text clustering task, which is like the number of concept clusters in Probase [26,30,10,28]. In the proposed **Co-Rank**: (i) we set the damping factor β_{CC} in Eq. (17) and β_{TT} in Eq. (21) and to 0.15 following the standard PageRank paradigm; (ii) we set the set of asymmetric parameters $\{\gamma_{CC}, \gamma_{CT}, \gamma_{TC}, \gamma_{TT}\}$ as 0.2,0.25, 0.35, 0.2, which yields the best results. (iii) we test the convergence threshold from $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$, and analysis results indicate that when the threshold tails off to below 10^{-5}, the co-ranking framework yields the best results in the short-text clustering task.

4.3 Experiments on Text Clustering

Because news title data, tweet data, Wiki first-sentence data and question data have no ground-truth labels, to evaluate the effectiveness of conceptualizing, we design text clustering task. Totally, we first generate "concepts"[3] of each short-text (in the aforementioned datasets) based on different algorithms, and then use these concepts as features to construct short-text vector, and run spherical K-means clustering [13,16] to evaluate each algorithm. This paper uses Purity [33], Adjusted Rand Index (ARI) [11] and Normalized Mutual Information (NMI) [25,4]to measure the quality of the short-text clustering task. The larger the Purity (ARI or NMI) is, the better the clustering result and the better the performance of the corresponding algorithm achieves.

We discuss these measurements as follows. Let $X = \{x_1, x_2, \cdots, x_{|X|}\}$ denote the set of short-text clusters after short-text clustering, wherein x_i indicates the i-th short-text cluster. Similarly, Let $Y = \{y_1, y_2, \cdots, y_{|Y|}\}$ denote the ground-truth set of short-text clusters. Besides, N denotes the total count of the short-text. Therefore, Purity could be measured as follows:

$$\text{Purity} = \frac{1}{N} \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \max |x_i \cap y_i| \tag{22}$$

Let n_{ij} denotes the count of short-texts which occurs in cluster x_i and cluster y_j simultaneously, and ARI could be defined as follows:

$$\text{ARI} = \frac{\sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \mathrm{C}_2^{n_{ij}} - \frac{\sum_{i=1}^{|X|} \mathrm{C}_2^{|x_i|} \cdot \sum_{j=1}^{|Y|} \mathrm{C}_2^{|y_j|}}{\mathrm{C}_2^N}}{\frac{\sum_{i=1}^{|X|} \mathrm{C}_2^{|x_i|} + \sum_{j=1}^{|Y|} \mathrm{C}_2^{|y_j|}}{2} - \frac{\sum_{i=1}^{|X|} \mathrm{C}_2^{|x_i|} \cdot \sum_{j=1}^{|Y|} \mathrm{C}_2^{|y_j|}}{\mathrm{C}_2^N}} \tag{23}$$

Moreover, the measurement of NMI is discussed as follows. Let $\mathrm{H}(X)$ denote the Information Entropy, defined as follows:

[3] Except for algorithm **LDA** generating topic as "concept", all the other algorithms generate concepts which are defined by lexical knowledge base Probase .

$$\mathrm{H}(X) = -\sum_{i=1}^{|X|} P(x_i) \cdot \log P(x_i) \tag{24}$$

Wherein, $P(x_i)$ indicates the probability that the short-text occurs in cluster x_i, and $P(y_j)$ indicates the probability that the short-text occurs in cluster y_j. Let $\mathrm{I}(X;Y)$ denotes the mutual information of set X and set Y, as follows:

$$\mathrm{I}(X;Y) = \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|}[P(x_i \cap y_j) \cdot \log \frac{P(x_i \cap y_j)}{P(x_i) \cdot P(y_j)}] \tag{25}$$

Therefore, we could utilize the following equation to define NMI:

$$\mathrm{NMI} = \frac{2 \cdot \mathrm{I}(X;Y)}{\mathrm{H}(X) + \mathrm{H}(Y)} \tag{26}$$

Table 1: Evaluation results of short-text clustering task.

	NewsTitle			**Twitter**			**WikiFirst**			**TREC**		
	Purity	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**
BOW	0.617	0.569	0.781	0.212	0.211	0.250	0.297	0.419	0.531	0.712	0.663	0.863
LDA	0.619	0.575	0.683	0.319	0.323	0.341	0.274	0.387	0.490	0.719	0.672	0.760
IJCAI$_{11}$	0.681	0.635	0.807	0.365	0.353	0.354	0.311	0.439	0.556	0.757	0.752	0.875
IJCAI$_{11}$+CL	0.711	0.651	0.809	0.378	0.351	0.387	0.326	0.460	0.583	0.812	0.748	0.881
IJCAI$_{15}$	0.737	0.675	0.822	0.419	0.381	0.416	0.343	0.484	0.613	0.832	0.770	0.882
RW	0.760	0.695	0.847	0.413	0.395	0.443	0.346	0.488	0.617	0.862	0.791	0.904
Co-Rank$_{AD}$	0.731	0.677	0.806	0.423	0.387	0.421	0.343	0.483	0.612	0.833	0.782	0.874
Co-Rank$_{IP}$	0.785	0.738	**0.879**	**0.461**	0.428	0.478	0.369	0.521	0.659	0.854	0.831	0.942
Co-Rank(Ours)	**0.801**	**0.753**	0.876	0.456	0.441	**0.482**	**0.380**	**0.537**	**0.679**	**0.871**	**0.848**	**0.961**

Experimental results are shown in Table 1. The results show the proposed co-ranking framework improves the baseline models in most cases: (i) **Co-Rank (Ours)** is superior to **Co-Rank$_{IP}$**, which ignores the the manual-defined knowledge (e.g., semantic similarity in lexical knowledge base); (ii) Taking measurement NMI as an example, **Co-Rank (Ours)** exceeds the recognized baseline model **IJCAI$_{15}$** by 6.57% and **IJCAI$_{11}$** by 8.55% on dataset **NewsTitle**, exceeds **IJCAI$_{15}$** by 15.87% and **IJCAI$_{11}$** by 36.16% on dataset **Twitter** (Note that, dataset **Twitter** is challenging because of its noise), and exceeds **IJCAI$_{15}$** by 8.94% and **IJCAI$_{11}$** by 22.08% on dataset **WikiFirst**, indicating that it is essential to utilize the beneficial interactions among terms and concepts.

5 Conclusion

Short-text conceptualization plays an increasingly vital role in text understanding and other applications. This paper proposes a novel co-ranking framework to address the

problem of short-text conceptualization, which operates an iterative procedure over a heterogeneous semantic graph and reinforces the terms and corresponding concepts simultaneously. Furthermore, this framework is found to automatically detect the contextual salient key-terms in the short-text. Experiments on real-world datasets suggest that the proposed co-ranking framework is effective.

Acknowledgements

The authors are very grateful to the editors and reviewers for their helpful comments. This work is funded by: (i) the China Postdoctoral Science Foundation (No.2018M641436); (ii) the Joint Advanced Research Foundation of China Electronics Technology Group Corporation (CETC) (No.6141B08010102); (iii) 2018 Culture and tourism think tank project (No.18ZK01); (iv) the New Generation of Artificial Intelligence Special Action Project (18116001); and (v) the Joint Advanced Research Foundation of China Electronics Technology Group Corporation (CETC) (No.6141B0801010a).

References

1. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Similarity search using concept graphs. In: 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 719–728 (2014)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: International Conference on World Wide Web. pp. 107–117 (1998)
4. Cho, D., Lee, B.: Optimized automatic sleep stage classification using the normalized mutual information feature selection (nmifs) method. Conf Proc IEEE Eng Med Biol Soc **2017**, 3094–3097 (2017)
5. Dathathri, R., Gill, G., Hoang, L., Dang, H.V., Brooks, A., Dryden, N., Snir, M., Pingali, K.: Gluon: a communication-optimizing substrate for distributed heterogeneous graph analytics. In: Acm Sigplan Conference on Programming Language Design & Implementation (2018)
6. Ding, Y.: Applying weighted pagerank to author citation networks. Journal of the Association for Information Science and Technology **62**(2), 236245 (2011)
7. Fujiwara, Y., Nakatsuji, M., Onizuka, M., Kitsuregawa, M.: Fast and exact top-k search for random walk with restart. Proceedings of the Vldb Endowment **5**(5), 442–453 (2012)
8. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research **34**(4), 443–498 (2014)
9. Hua, W., Wang, Z., Wang, H., Zheng, K., Zhou, X.: Short text understanding through lexical-semantic analysis. In: IEEE International Conference on Data Engineering. pp. 495–506 (2015)
10. Huang, H., Wang, Y., Feng, C., Liu, Z., Zhou, Q.: Leveraging conceptualization for short-text embedding. IEEE Transactions on Knowledge and Data Engineering **30**(7), 1282–1295 (2018)
11. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (1985)
12. Kim, D., Wang, H., Oh, A.: Context-dependent conceptualization. In: International Joint Conference on Artificial Intelligence. pp. 2654–2661 (2013)
13. Li, M., Xu, D., Zhang, D., Zou, J.: The seeding algorithms for spherical k -means clustering. Journal of Global Optimization pp. 1–14 (2019)

14. Li, X., Roth, D.: Learning question classifiers. In: 19th International Conference on Computational linguistics. pp. 1–7 (2002)
15. MauroGasparini: Markov chain monte carlo in practice. Technometrics **39**(3), 338–338 (1999)
16. Peterson, A.D., Ghosh, A.P., Maitra, R.: Merging k-means with hierarchical clustering for identifying general-shaped groups. Stat **7**(1) (2018)
17. Rui, Y., Lapata, M., Li, X.: Tweet recommendation with graph co-ranking. In: Meeting of the Association for Computational Linguistics: Long Papers (2012)
18. Salton, G., Mcgill, M.J.: Introduction to modern information retrieval. McGraw-Hill (1983)
19. Sebastian, Y., Eu-Gene, S., Orimaye, S.O.: Learning the heterogeneous bibliographic information network for literature-based discovery. Knowledge-Based Systems **115**, 66–79 (2016)
20. Shi, C., Li, Y., Zhang, J., Sun, Y., Yu, P.S.: A survey of heterogeneous information network analysis. IEEE Trans. Knowl. Data Eng. **29**, 17–37 (2017)
21. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: International Joint Conference on Artificial Intelligence. pp. 2330–2336 (2011)
22. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three. pp. 2330–2336 (2011)
23. Song, Y., Wang, S., Wang, H.: Open domain short text conceptualization: a generative + descriptive modeling approach. In: International Conference on Artificial Intelligence. pp. 3820–3826 (2015)
24. Song, Y., Wang, S., Wang, H.: Open domain short text conceptualization: a generative + descriptive modeling approach. In: Proceedings of the 24th International Conference on Artificial Intelligence (2015)
25. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions (2003)
26. Wang, F., Wang, Z., Li, Z., Wen, J.R.: Concept-based short text classification and ranking. In: The ACM International Conference. pp. 1069–1078 (2014)
27. Wang, Y., Huang, H., Feng, C.: Query expansion based on a feedback concept model for microblog retrieval. In: International Conference on World Wide Web. pp. 559–568 (2017)
28. Wang, Y., Huang, H., Feng, C., Zhou, Q., Gu, J., Gao, X.: Cse: Conceptual sentence embeddings based on attention model. In: 54th Annual Meeting of the Association for Computational Linguistics. pp. 505–515 (2016)
29. Wang, Z., Cheng, J., Wang, H., Wen, J.: Short text understanding: A survey. Journal of Computer Research and Development **53**(2), 262–269 (2016)
30. Wang, Z., Zhao, K., Wang, H., Meng, X., Wen, J.R.: Query understanding through knowledge-based conceptualization. In: International Conference on Artificial Intelligence. pp. 3264–3270 (2015)
31. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: SIGMOD Conference (2012)
32. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase:a probabilistic taxonomy for text understanding. In: ACM SIGMOD International Conference on Management of Data. pp. 481–492 (2012)
33. Ying, Z., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning **55**(3), 311–331 (2004)
34. Zhou, D., Orshanskiy, S.A., Zha, H., Giles, C.L.: Co-ranking authors and documents in a heterogeneous network. In: 7th IEEE International Conference on Data Mining. pp. 739–744 (2007)