A Fast and Effective Method for Clustering Large-Scale Chinese Question Dataset

Xiaodong Zhang Houfeng Wang

Institute of Computational Linguistics, Peking University

December 9, 2014

Outline

Introduction

- Community Question Answering (CQA) websites have accumulated large archives of question-answer pairs
- Application: QA system
- Question retrieval in large dataset is slow
- A feasible way: clustering

Difficulties of question clustering

- Vector Space Model
  - Data sparseness
  - Lexical gap
- Embedding methods
  - Not interpretable

Example:

- Question 1: 电脑出故障，过了保修期怎么办？
  (My computer broke down and its warranty expired. What should I do?)
- Question 2: 我想给笔记本装个固态硬盘，哪个牌子比较好？
  (I would like to install a SSD to my laptop. Which brand is good?)

Question Similarity Measure

- Semantics is introduced into VSM
- Word relatedness calculated by word2vec with cosine similarity

$$r(t_1, t_2) = \begin{cases} 1 & t_1 = t_2 \\ 0 & t_1 \text{ or } t_2 \text{ not in training data} \\ \dfrac{\langle v_1 \cdot v_2 \rangle}{\|v_1\| \, \|v_2\|} & \text{otherwise} \end{cases}$$

- The similarity of two questions are modeled by a bipartite graph $G = (U, V, E)$

Construct the bipartite graph

- Construct the bipartite graph:
  For each word type t_{1i} ($i \in [1, n_1]$, n_1 is the number of word types in q_1) in q_1, there is a corresponding node u_i in U. Node v_j in V for each word type t_{2j} in q_2 is defined in the same way. If the relatedness of t_{1i} and t_{2j} exceeds a threshold, the two words are considered related and nodes u_i and v_j are connected by a edge, of which the weight is calculated as follows:

$$e_{ij} = w_{1i} \times w_{2j} \times r_{ij}$$

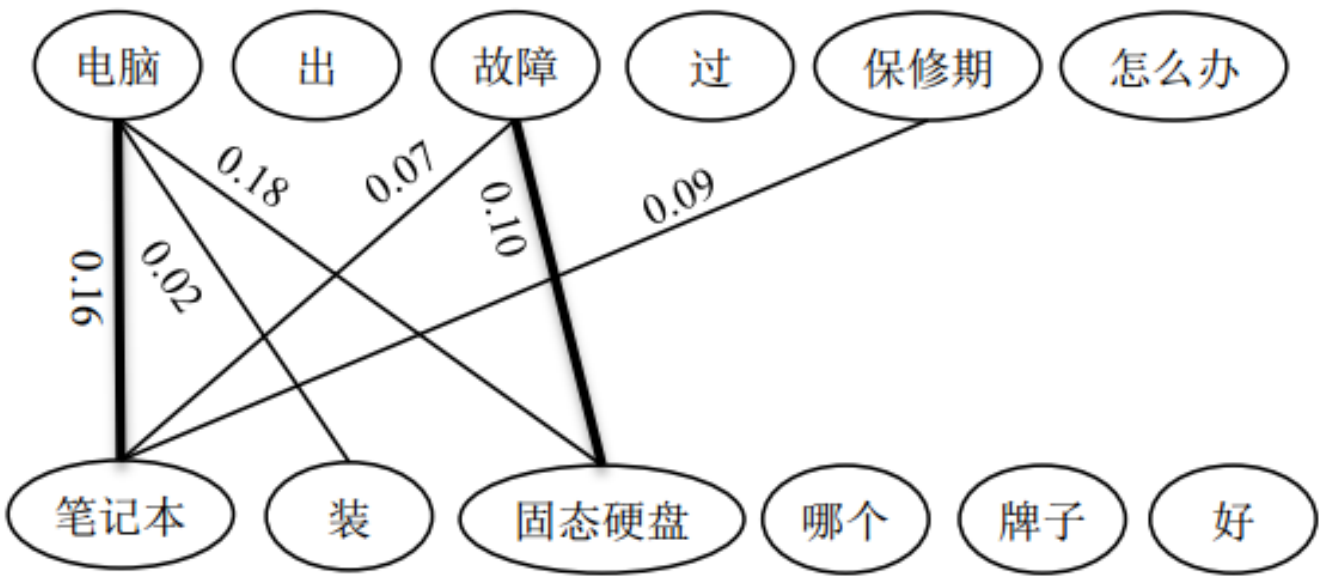
- The similarity of two questions is defined as the sum of weights of edges in the maximum weight matching of the bipartite graph. Formally,

$$s = \sum_{e \in \mathrm{MWM}(G)} w_e$$

Example

- Question vectors:
  - v1: [电脑/0.44, 出/0.11, 故障/0.55, 过/0.17, 保修期/0.67, 怎么办/0.11];
  - v2: [笔记本/0.47, 装/0.27, 固态硬盘/0.80, 哪个/0.07, 牌子/0.33 好/0.07].
- Pairs of words that relatedness exceeds 0.2 (threshold):

  (电脑, 笔记本) = 0.78; (电脑, 固态硬盘) = 0.52; (电脑, 装) = 0.21; (故障, 笔记本) = 0.28; (故障, 固态硬盘) = 0.22; (保修期, 笔记本) = 0.28.
- The similarity of the two questions are 0.26



Semantic K-means Algorithm

Algorithm 1 semantic k-means (k, D)

```
1: choose k data points randomly as the initial centroids (cluster centers);
2: repeat
3:     for each data point x ∈ D do
4:         if in the first iteration then
5:             compute the similarity of x and each centroid by our proposed similarity mea-
               sure;
6:         else
7:             compute the similarity of x and each centroid by cosine similarity
8:         end if
9:         assign x to the most similar centroid
10:    end for
11:    re-compute the centroid using the current cluster memberships
12: until the stopping criterion is met
```

Extended Semantic K-means Algorithm

Algorithm 2 extended semantic k-means (k, D, m, d)

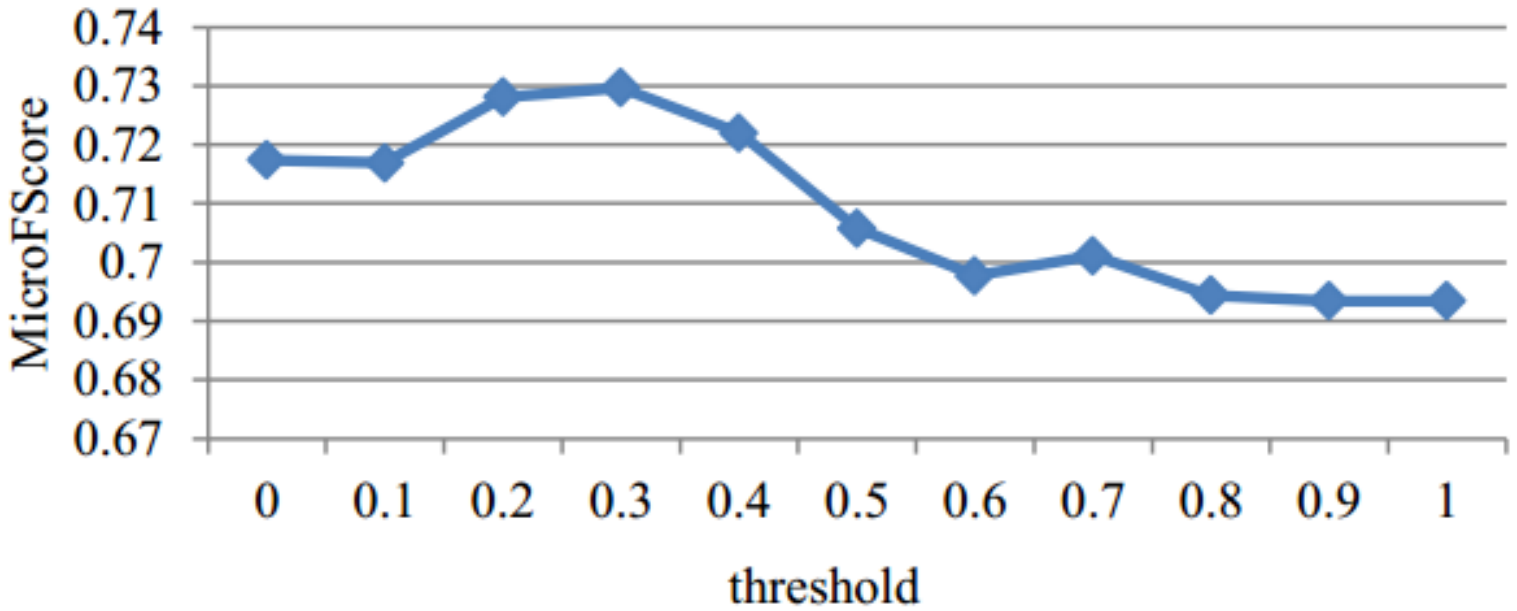
choose k data points randomly as the initial centroids (cluster centers);
$iter \leftarrow 1$
repeat
    truncate and normalize centroids (reserve d dimensions)
    for each data point $x \in D$ **do**
        compute the similarity of x and each centroid by our proposed similarity measure;
        assign x to the most similar centroid
    end for
    re-compute the centroid using the current cluster memberships
    $iter \leftarrow iter + 1$
until $iter > m$
repeat
    for each data point $x \in D$ **do**
        compute the similarity of x and each centroid by cosine similarity;
        assign x to the most similar centroid
    end for
    re-compute the centroid using the current cluster memberships
until the stopping criterion is met

Experiments

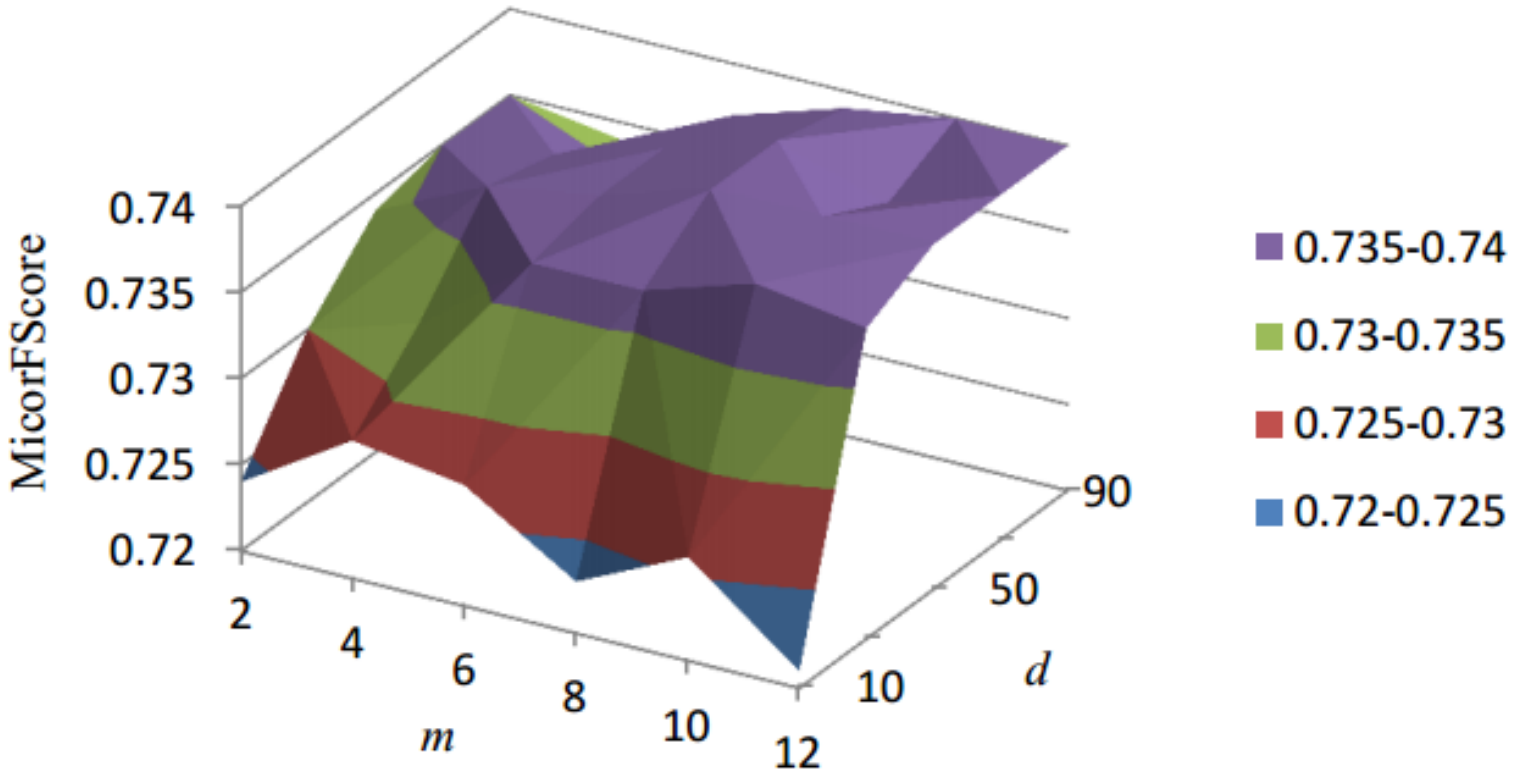
- Dataset:
  16000 questions, 8 class
- Results:

#	Method	MicroFScore (average)	MicroFScore (highest)	Time
1	k-means	0.644	0.751	6s
2	spectral	0.554	0.575	919s
3	spectral++	0.671	0.721	1725s
4	wiki	0.678	0.757	207s
5	LDA	0.734	0.798	32s
6	BTM	0.741	0.804	148s
7	Sk-means	0.736	0.821	10s
8	ESk-means	0.740	0.821	88s

Relatedness Threshold



Parameter m and d in ESk-means



Conclusion

- A novel similarity measure for questions
  - Word relatedness
  - bipartite graph
- Question clustering methods
  - Sk-means
  - ESk-means
- Comparisons between our methods and some mainstream methods
- Future works
  - Explore the application of our similarity measure in other clustering methods and NLP tasks

Thank you!