Hierarchical Conceptual Labeling

Haiyun Jiang[1], Cengguang Zhang[1], Deqing Yang[1], Yanghua Xiao[1,2⋆], Jingping Liu[1], Jindong Chen[1], Chao Wang[1], Chenguang Li[1], Jiaqing Liang[1], Bin Liang[1], and Wei Wang[1]

[1] School of Computer Science, Fudan University, Shanghai, China
[2] Shanghai Institute of Intelligent Electronics & Systems, Shanghai, China
{jianghy16, cgzhang15, yangdeqing, shawyh, liujp17, chenjd17, cwang17, cgli17, liangbin, weiwang1}@fudan.edu.cn {l.j.q.light}@gmail.com

Abstract. The bag-of-words model is widely used in many AI applications. In this paper, we propose the task of hierarchical conceptual labeling (HCL), which aims to generate a set of conceptual labels with a hierarchy to represent the semantics of a bag of words. To achieve it, we first propose a denoising algorithm to filter out the noise in a bag of words in advance. Then the hierarchical conceptual labels are generated for a clean word bag based on the clustering algorithm of Bayesian rose tree. The experiments demonstrate the high performance of our proposed framework.

1 Introduction

The bag-of-words model is widely used in many natural language processing tasks. There are lots of mature technologies to generate a bag of words (BoW) [4]. However, a BoW is just a collection of scattered words and it is difficult to be understood by machines or human beings without explicit semantic analysis. The conceptualization-based methods, i.e., *conceptual labeling* (CL), aim to generate conceptual labels for a BoW to explicitly represent its semantics. In [6, 5, 3], a BoW is first divided into multiple groups according to their semantic relevance and then each group is labeled with a concept that can specifically summarize the explicit semantics. We present two examples as follows.

In this paper we propose the task of *hierarchical conceptual labeling* (H-CL), which represents the semantics of a BoW by *hierarchical* conceptual labels (i.e., a label set with different granularities). For example, given a BoW {China,Japan,France,Germany,Russia}, the hierarchical conceptual labels can be {*Asian country, EU State*} and {*country*}. In general, the hierarchical labels contain more information, which allows real applications to select labels with different abstractness according to their real requirements.

We consider the hierarchical cluster algorithm: Bayesian rose tree (BRT) [1] as our framework to generate hierarchical conceptual labels, where the candidate concepts are derived from the knowledge base: Microsoft concept graph (MCG) [8]. Besides, we also propose a simple but effective method to delete the noise before the conceptualization operation.

[⋆] Correspondent author. This paper was supported by National Natural Science Foundation of China under No.61732004.

2 Framework

We first present how to filter out the noise words in a BoW. Then we elaborate on the generation process of hierarchical conceptual labels.

2.1 Filtering Out Noise

The basic idea is: *if a word in a BoW is hard to be semantically clustered with any other word, i.e., difficult to be tagged with the same conceptual label as any other word, then we take it as noise and remove it from the BoW.*

Specifically, let $\mathcal{D}$ be the input BoW, and d_i (d_j) be the i-th (j-th) instance[3] in $\mathcal{D}$. We take $p(c|d_i, d_j)$ to measure how well the concept c conceptualizes the semantics of two instances d_i, d_j. We use Bayesian rule to compute $p(c|d_i, d_j)$ as follows:

$$p(c|d_i, d_j) = \frac{p(d_i, d_j|c)p(c)}{p(d_i, d_j)} = \frac{p(d_i|c)p(d_j|c)p(c)}{p(d_i)p(d_j)} \tag{1}$$

Then $p(c|d_i, d_j) = \frac{1}{\tilde{p}^2} p(d_i|c)p(d_j|c)p(c)$. The prior probability $p(c)$ measures the popularity of c. Intuitively, a larger $p(c|d_i, d_j)$ indicates c can summarize d_i and d_j well, so d_i and d_j have strong semantic relevance. $p(d_k|c)$ and $p(c)$ are estimated using knowledge in MCG [8]. Let $\mathcal{C}_i$ and $\mathcal{C}_j$ be the concept sets of d_i and d_j in MCG, respectively. $\mathcal{C}_{i,j} = \mathcal{C}_i \cap \mathcal{C}_j$ denotes the *shared concept set* of d_i and d_j. We describe the denoising algorithm as follows.

Consider the word $d_i \in \mathcal{D}$, for any other word $d_j \in \mathcal{D}$ ($d_j \neq d_i$), if we cannot find an appropriate concept in $\mathcal{C}_{i,j}$ to conceptualize d_i and d_j, i.e.,

$$\max_{d_j \in \mathcal{D}, c \in \mathcal{C}_{i,j}} p(c|d_i, d_j) < \delta \tag{2}$$

then d_i is treated as noise. δ is a hyperparameter.

2.2 Hierarchical Conceptual Labeling

Next, we describe how to generate hierarchical conceptual labels for a BoW. The basic idea is: *clustering a BoW $\mathcal{D}$ hierarchically based on BRT [1], and for each cluster $\mathcal{D}_m$ an appropriate conceptual label will be generated.* We present the pseudo code in Algorithm 1.

Estimation of $\boldsymbol{f(\mathcal{D}_m)}$ and $\boldsymbol{p(D_m|T_m)}$. $f(\mathcal{D}_m)$ qualifies the probability that all the words in $\mathcal{D}_m$ belong to the same cluster and it further helps us to estimate $p(D_m|T_m)$. Similar to [7], we consider that $\mathcal{D}_m$ with more shared concepts in MCG is more inclined to belong to the same cluster. For each $c \in \mathcal{C}_m$ (the shared concepts of $\mathcal{D}_m$), the probability that $\mathcal{D}_m$ belongs to the same cluster is computed as

$$p(\mathcal{D}_m|ct.) = \prod_{d_i \in \mathcal{D}_m} p(d_i|c.) \tag{5}$$

When considering all the concepts in $\mathcal{C}_m$, $f(\mathcal{D}_m)$ is computed by $f(\mathcal{D}_m) = \sum_{c \in \mathcal{C}_m} p\left(c\right) p\left(\mathcal{D}_m \left| c\right.\right)$. Based on $f(\mathcal{D}_m)$, the probability $p(D_m|T_m)$ can be recursively calculated by $p\left(\mathcal{D}_m \left| T_m\right.\right) = \pi_m f(\mathcal{D}_m) + (1 - \pi_m) \prod_{T_k \in \mathrm{ch}(T_m)} p\left(D_k \left| T_k\right.\right)$.

[3] In this paper, the words in BoWs are also called instances.

Algorithm 1 Hierarchical conceptual labeling based on the Bayesian rose tree.

```
Input: data $\mathcal{D} = \{d_1, d_2, \cdots, d_N\}$
Output: hierarchical conceptual labels
Initialize: LabelSet= {}, number of clusters $k = N$, $\mathcal{D}_i = \{d_i\}$, $T_i = \{d_i\}$, $p(\mathcal{D}_i|T_i) = 1$ $(i = 1, \cdots, N)$ and $L(T_m) = \gamma_0$
while $k > 1$ and $L(T_m) > \gamma$ do
    Find the pair of trees $T_i$ and $T_j$ and the merge operation that can maximize the likelihood ratio:
        $L(T_m) = \frac{p(\mathcal{D}_m|T_m)}{p(\mathcal{D}_i|T_i)\, p(\mathcal{D}_j|T_j)}$    (3)
    Select the conceptual label $c_m^*$:
        $c_m^* = \arg\max_{c \in \mathcal{C}_m} p(c|\mathcal{D}_m)$    (4)
    Merge $T_i$ and $T_j$ into $T_m$ by the selected merge operation; $\mathcal{D}_m \leftarrow \mathcal{D}_i \cup \mathcal{D}_j$; Add $c_m$ to LabelSet; Delete $T_i$ and $T_j$, $k \leftarrow k - 1$
end while
```

Estimation of $\boldsymbol{\pi_m}$. π_m is a hyperparameter denoting the prior probability that the leaves under T_m are kept in one cluster rather than subdivided by the recursive partitioning process. We simply set $\pi_m = 0.5$ in this paper.

Label Generation. To generate hierarchical conceptual labels for a BoW, we need to select an appropriate conceptual label to well conceptualize each cluster $\mathcal{D}_m$. The following criterion is used to select the most appropriate conceptual label:

$$c_m^* = \arg\max_{c \in \mathcal{C}_m} p(c|\mathcal{D}_m) = \arg\max_{c \in \mathcal{C}_m} p(\mathcal{D}_m|c)p(c) \tag{6}$$

Likelihood ratio $\boldsymbol{\gamma}$. In most cases, a BoW is hard to be semantically merged into *one* cluster, so the cluster operation should be stopped when there is no appropriate label. We take a likelihood ratio γ, and stop clustering when $L(T_m) < \gamma$.

3 Experiments

We evaluate the generated hierarchical conceptual labels. In all experiments, $\delta = 5 \times 10^{-8}$ and $\gamma = 0.8$ are used.

Dataset. The dataset in [7] is used, which contains two subsets: Flickr and Wikipedia. We sample $b = 500$ BoWs from each dataset for evaluation.

Baselines. To the best of our knowledge, there is no work to deal with the task of HCL, so we present two strong baselines constructed by ourselves. *(1) Bayesian hierarchical clustering-based model (BHC).* We first cluster a BoW using Bayesian hierarchical clustering [2]. Each node in the hierarchy is equipped with a concept to conceptualize the corresponding cluster, where the candidate concepts are also from MCG. *(2) Maximal clique segmentation-based model (MC-S).* We first construct a semantic graph for a BoW, where the vertex corresponds to a word. Then we take the maximal clique segmentation [5] to split the graph into several subgraphs given a similarity threshold. Finally, we select one conceptual label for each graph, thus generating a flat conceptual label set for a BoW. Furthermore, when considering multiple similarity thresholds, we will get the multiple label sets with different granularities for a BoW.

Metric. We evaluate the models and consider the two cases: with (without) denoising algorithm. We recruit $v = 5$ volunteers to evaluate the labeling results by scoring ($0 \leq score \leq 3$), where the scoring criteria are motivated by [7]. The average score is computed by $\frac{1}{bv}\sum_{i=1}^{v}\sum_{j=1}^{b} s_{i,j}$, where $s_{i,j}$ is the score of the j-th BoW by volunteer i, b is number of BoWs in each dataset and v is the number of volunteers.

Table 1. Average scores on Flickr and Wikipedia data.

Model	Flickr	Wikipedia	Model	Flickr	Wikipedia
BHC	0.228	0.233	BHC + Denoising	0.247	0.261
MCS	0.240	0.245	MCS + Denoising	0.266	0.271
BRT	**0.251**	**0.264**	BRT + Denoising	**0.273**	**0.282**

Results and Analysis. The results are presented in Table 1. We conclude that (1) the scores with the denoising algorithm are higher than these without it for all models, which proves the effectiveness of the denoising method. (2) The proposed model outperforms the other two baselines. In particular, BHC only considers the binary branching structures in the hierarchy and cannot generate multi-branching structures that frequently appear in the BoW clustering. MCS only clusters BoWs into multi-level label sets without hierarchy.

4 Conclusion

This paper first proposes the task of HCL, which aims to generate conceptual labels with different granularities for BoWs. To achieve it, we propose the BRT-based approach with high performance. Besides, we also propose a denoising algorithm to effectively filter out the noise in advance.

References

1. Blundell, C., Teh, Y.W., Heller, K.A.: Bayesian rose trees. In UAI (2010)
2. Heller, K.A., Ghahramani, Z.: Bayesian hierarchical clustering. In ICML 21 (2005)
3. Hua, W., Wang, Z., Wang, H., Zheng, K.: Short text understanding through lexical-semantic analysis. In: IEEE International Conference on Data Engineering. pp. 495–506 (2015)
4. Pay, T.: Totally automated keyword extraction. 2016 IEEE International Conference on Big Data (Big Data) pp. 3859–3863 (2016)
5. Song, Y., Wang, H., Wang, H.: Open domain short text conceptualization: a generative + descriptive modeling approach. In: International Conference on Artificial Intelligence. pp. 3820–3826 (2015)
6. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledge base. IJCAI pp. 2330–2336 (2011)
7. Sun, X., Xiao, Y., Wangy, H., Wang, W.: On conceptual labeling of a bag of words. IJCAI pp. 1326–1332 (2015)
8. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: A probabilistic taxonomy for text understanding. In SIGMOD pp. 481–492 (2012)