Word Clustering Based on Un-LP Algorithm

Jiguang Liang[1], **Xiaofei Zhou**[1], **Yue Hu**[1], **Li Guo**[1*], **Shuo Bai**[1,2]
[1]National Engineering Laboratory for Information Security Technologies,
Institute of Information Engineering, Chinese Academy of Sciences,
Beijing 100190, China
[2]Shanghai Stock Exchange, Shanghai 200120, China
{liangjiguang, zhouxiaofei, huyue, guoli, baishuo}@iie.ac.cn

Abstract

Word clustering which generalizes specific features cluster words in the same syntactic or semantic categories into a group. It is an effective approach to reduce feature dimensionality and feature sparseness which are clearly useful for many NLP applications. This paper proposes an unsupervised label propagation algorithm (Un-LP) for word clustering which uses multi-exemplars to represent a cluster. Experiments on a synthetic 2D dataset show the strong ability of self-correcting of the proposed algorithm. Besides, the experimental results on 20NG demonstrate that our algorithm outperforms the conventional cluster algorithms.

1 Introduction

Word clustering is the task of the division of words into a certain number of clusters (groups or categories). Each cluster is required to consist of words that are similar to one another in syntactic or semantic construct and dissimilar to words in distinctive groups. Word clustering generalizes specific features by considering the common characteristics and ignoring the specific characteristics among the individual features. It is an effective approach to reduce feature dimensionality and feature sparseness (Han et al., 2005).

Recently, word clustering offers great potential for various useful NLP applications. Several studies have addressed dependency parsing (Koo et al., 2008; Sagae and Gordon, 2009). Momtazi and Klakow (2009) propose a word clustering approach to improve the performance of sentence retrieval in Question Answering (QA) systems. Wu et al. (2010) present an approach to integrate word clustering information into the process of unsupervised feature selection. Sun et al. (2011) use large-scale word clustering for a semi-supervised relation extraction system. It also contributes to word sense disambiguation (Jin et al., 2007), named entity recognition (Turian et al., 2010), part-of-speech tagging (Candito and Seddah, 2010) and machine translation (Uszkoreit and Brants, 2008; Jeff et al., 2011).

This paper presents an unsupervised algorithm for word clustering based on a probabilistic transition matrix. Given a text document dataset, a list of words is generated by removing stop words and very unfrequent words. Each word is required to be represented by the documents in the dataset, which results in a co-occurrence matrix. By calculating the similarity of words, a word similarity graph with transition (propagation) probabilities as weight edges is created. Then, a new kind word clustering algorithm, based on label propagation, is applied.

The remaining parts of this paper are organized as follows: Section 2 formulates word clustering problem in the context of unsupervised learning. Then we describe the word clustering algorithm in Section 3 and present our experiments in Section 4. Finally we conclude our work in Section 5.

2 Problem setup

Assume that we have a corpus with N documents denoted by $D = \{d_1, d_2, \cdots, d_N\}$; each document in the corpus consists of a list of words denoted by $d_i = \{w_1, w_2, \cdots, w_{N_d}\}$ where each w_i is an item from a vocabulary index with V distinct terms denoted by $W = \{v_1, v_2, \cdots, v_V\}$ and N_d is the document

Proceedings of the AHA! Workshop on Information Discovery in Text, pages 25–30,
Dublin, Ireland, August 23 2014.

Algorithm 1 Semi-supervised LP Algorithm

```
Input:
$W_l = \{v_i\}_{i=1}^{l}$ labeled data
$W_u = \{v_i\}_{i=u}^{V}$ unlabeled data
$\overline{T} = \{T_{ij}\}\ 1 \leq i, j \leq V$ transition matrix
Output: $Y_U$
Begin
    Row-normalize $T$ by $\overline{T}_{ij} = T_{ij} / \sum_{k=1}^{V} T_{ik}$
    While not converged do
        Propagate the labels by $Y^{t+1} = \overline{T} Y^t$
        Row-normalize $Y^{t+1}$
        Clamp the labeled data
    End while
End
Return $Y_U$
```

Algorithm 2 Unsupervised LP Word Clustering

```
Input:
$W = \{v_i\}_{i=1}^{u}\ (u = V)$ unlabeled words
$\overline{T}_{uu} = \{T_{ij}\}\ 1 \leq i, j \leq V$ transition matrix
Output:
$\Lambda = \{(\Lambda_1, \Lambda_2, \cdots, \Lambda_L\}$ word-clusters
Begin
    $\{V_L^0, Y_L, \overline{T}_{ul}^{\,0}\} = initialization(W)$
    While not converged do
        $Y_U^{t+1} = Semi - LP(V_L^t, Y_L^t, \overline{T}_{ul}^0, \overline{T}_{uu})$
        $\Lambda^{t+1} = partition\_cluster(Y_U^{t+1})$
        $\{V_L^{t+1}, \overline{T}_{ul}^{\,t+1}\} = update(\Lambda^{t+1})$
    End while
End
Return $\Lambda^{t+1}$
```

length. We define the vector of word v_i in the vocabulary to be $v_i =< v_{id_1}, v_{id_2}, \cdots, v_{id_N} >$. This allows us to define a $V \times N$ word-document matrix WD for the vocabularies. WD_{ij} is equal to 1 if $v_i \in d_j$ and equal to 0 otherwise. Then we take these words as the vertices of a connected graph. In this paper, we define the edge weight ω_{ij} as the co-occurrence frequency between v_i and v_j. Obviously, we expect that larger edge weights allow labels to travel through more easily. So we define a $V \times V$ probabilistic transition matrix T where $T_{ij} = P(v_j \rightarrow v_i) = \omega_{ij} / \sum_{k=1}^{V} \omega_{kj}$.

The L value which is used to represent the number of word clusters is specified. We define a $V \times L$ label matrix Y. Clearly, $y_i \in Y$ represents the label probability distributions of word v_i and $Y_i^* = argmax\ Y_{ik}(0 < k \leq L)$ is its cluster label. For example, suppose $L = 3$ and a word v has a label distribution $y =< 0.1, 0.8, 0.1 >$, it implies that v belongs to the second class.

3 Unsupervised LP Word Clustering

Label propagation (Zhu and Ghahramani, 2002) is a semi-supervised algorithm (Semi-LP) which needs labeled data. Let $\{(v_1, y_1), \cdots, (v_l, y_l)\}$ be labeled data, $\{(v_{l+1}, y_{l+1}), \cdots, (v_{l+u}, y_{l+u})\}$ be unlabeled ones where $l + u = V$, $Y_L = [y_1, , \cdots, y_l]^T$ and $Y_U = [y_{l+1}, \cdots, y_{l+u}]^T$. Y_U is un-known and $l << u$. The label propagation algorithm is summarized in Algorithm 1.

In clustering problems, the goal is to select a set of exemplars from a dataset that are representative of the dataset and each cluster is represented by one and only one exemplar (Krause and Gomes, 2010). However, these exemplars are just all Semi-LP needs for clustering. LP lacks labeled data when is used for unsupervised learning. In this paper, we are interested in partitioning words into several clusters without any label priori using unsupervised LP (Un-LP) algorithm. Firstly we randomly select K ($K \geq L$, usually K is a multiple of L) words to construct an exemplar set $E = \{E_i\}_{i=1}^{K}$ which is different from the conventional exemplar-based cluster algorithms, assign class labels to them and construct the corresponding probabilistic transition matrix $\overline{T}_{ul}^{\,0}$ ($initialization$). These exemplars are considered as labeled words and the rest $U = W - E$ are unlabeled words. $\overline{T}_{ul}$ is the probability of transition from unlabeled words to labeled ones. At this step, it needs the assurance that each class could be represented by at least one exemplar and each exemplar could only be assigned one class label.

Now the connected weighted graph consists of two parts: $G = (E \cup U, \overline{T}_{ul} \cup \overline{T}_{uu})$ where $\overline{T}_{uu}$ is the transition probability between unlabeled words. Next, our algorithm iterates between the following three steps: given a set of LP parameters, we first propagate labels to unlabeled words with the initial label distributions and get the corresponding labels ($Semi - LP$). Then, these derived label distributions are used to guide the partitioning of unlabeled data ($partition\_cluster$) to L clusters. We use residual sum of squares (RSS) to choose the most centrally located words and replace the old exemplars that represent the cluster. Specifically, for a word cluster $c_i = \{v_1, \cdots v_n\}$, $RSS_i = \sum_{j=1}^{n} \omega_{ij}$. Then we sort RSS_i $(0 < i < n)$ and update exemplars by the words with bigger RSS for this cluster ($update$). All of the above steps, summarized in Algorithm 2, are iterated until convergence.

4 Experiment

4.1 Experimental Setup

To demonstrate properties of our proposed algorithm we investigate both a synthetic dataset and a real-world dataset. Figure 1(a) shows the synthetic dataset. For a real world example we test Un-LP on a subset of 20 Newsgroups (20NG) dataset which is preprocessed by removing common stop-words and stemming. We use the classes $atheism$, $hardware$, $hockey$ and $space$ for test and randomly select 300 samples from each class as the test dataset in this section. However, 20NG is not suited for word clustering evaluation. So, firstly, we reconstruct it by pair-wise testing which is a specification-based testing criterion. Then we can obtain six ($C_4^2 = 6$) pairwise subsets represented by $\{D_1, \cdots, D_6\}$. In order to facilitate the evaluation, we use those words that only occur in one class for clustering.

4.2 Exemplar Self-correction

This multi-step iterative method is simple to implement and surprisingly effective even with wrong initial labeled data. To illustrate the point, we describe a simulated dataset with two-moon pattern. Obviously, the points in one moon should be more similar to each other than the points across the moons as shown in Figure 1(b). During the initialization phase, four points in the lower moon are selected and assigned with different labels. The exemplars of the upper moon are mis-labeled as shown in Figure 1(c). In the next five iteration steps, exemplars have been gradually moved to the center of the upper moon. Finally, when $t \geq 5$ Un-LP converges to a fixed assignment, which achieves an ideal cluster result.

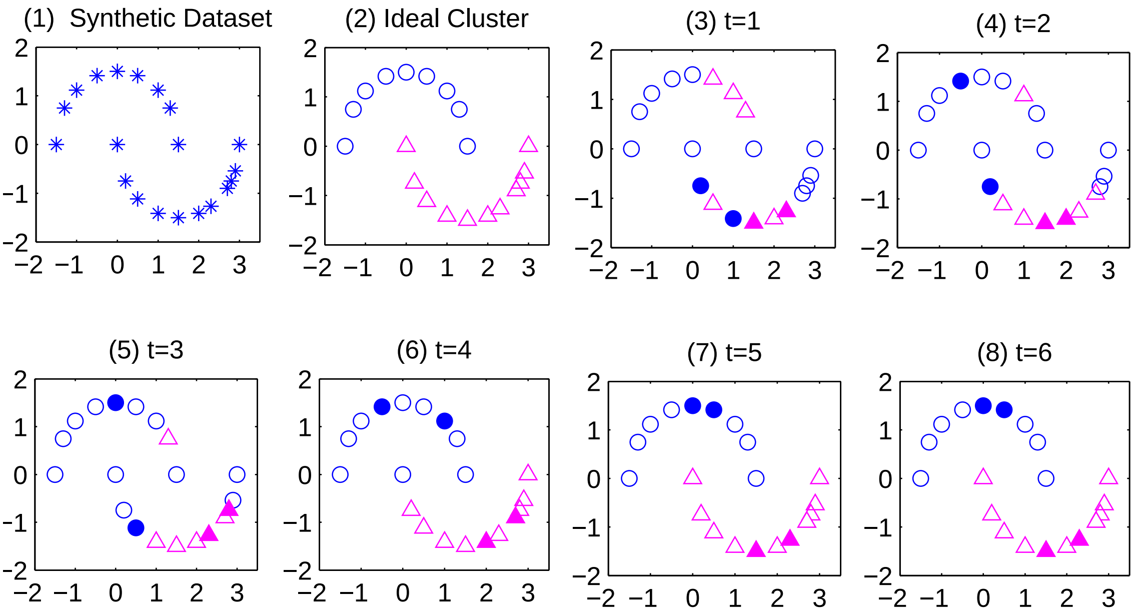

Figure 1: Clustering result of unsupervised LP clustering algorithm on two-moon pattern dataset. (a) Two-moon pattern dataset without any labeled points, (b) ideal clustering results. The convergence process of unsupervised LP with t from 1 to 6 is shown from (c) to (h). Solid points are labeled data that are selected to represent the clusters.

4.3 Word Clustering Performance

This section provides empirical evidence that the proposed algorithm performs well in the problem of word clustering. Figure 2 shows the mean precisions and recalls over 10 runs of the baseline algorithms as well as Un-LP.

From Figure 2, it can be clearly observed that Un-LP ($K/L = 5$) yields the best performance, followed by Semi-LP with 20 labeled words. In general, the recalls with k-means and k-medoids are higher, while the precisions are much lower. Figure 2 also shows the results of other two semi-supervised word

Cluster1	Cluster2	Cluster3	Cluster4
Atheism	Hardware	Hockey	Space
geode **religiously** **benefactor** *meng* stacker *mcl* **mormon** madden **mythology** *timmons* *cbnewsj* **agnostics** *fanatism* *engr* *chade* *tan* *falsifiable* *existed* *ucsb* *sentence*	**driver** **soundblaster** **cards** isbn **manufacturer** portal prize **mastering** **connectors** **floppies** dock **adapter** **multimedia** **installing** bowman *configure* *physchem* *jumpers* **motherboardsfdisk** **seagate**	**goalies** bug *hfd* *johansson* breton **scorers** *carpenter* *stevens* *smythe* *janney* *fleury* *vancouver* *stl* *cheveldae* *selanne* *winnipeg* *canadiens* *bure* *nyr* *capitals*	hub **atom** *aug* *larson* *sts* **orbital** skydive parity **accelerations** *desire* *anniversary* *projects* digital *protection* atari **temperatures** *voyagers* *zoology* *updated* *teflon*

Table 1: Top-20 words extracted by unsupervised LP word cluster algorithm.

clustering algorithms, PCK-means (Basu et al., 2004) and MPCK-means (Bilenko et al., 2004) with 200 must-link and cannot-link constraints. Also when comparing these unsupervised and semi-supervised approaches previously mentioned, we can find that our unsupervised algorithm consistently achieves significantly better results. Therefore, unsupervised LP seems to be a more reasonable algorithm design in terms of word clustering.

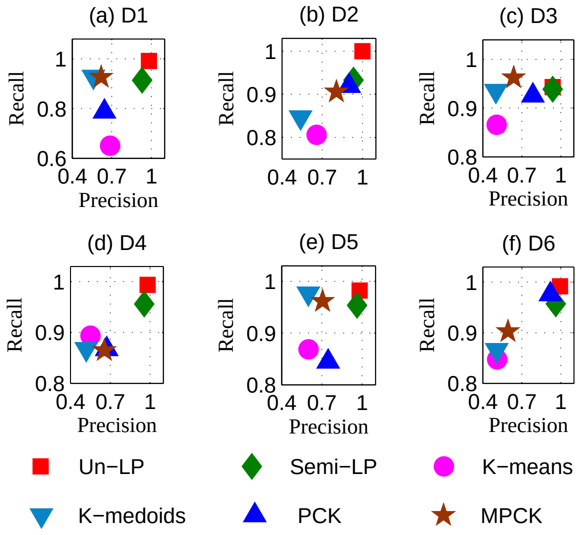

Figure 2: Precision vs. recall of clustering results on 20NG where D_1 = $\{atheism$ vs. $hardware\}$, D_2 = $\{atheism$ vs. $hockey\}$, D_3 = $\{atheism$ vs. $space\}$, D_4 = $\{hardware$ vs. $hcokey\}$, D_5 = $\{hardware$ vs. $space\}$ and D_6 = $\{hockey$ vs. $space\}$.

4.4 Effect of exemplar number e

We now investigate how the number of exemplar (e) for each cluster affects the clustering. In particular, we are interested in performance under conditions when the number of exemplar grows - which is the motivation for using more than one exemplars to represent a cluster. From Figure 3, we can observe that when more words are labeled, Semi-LP shows further improvement in F-value. However, the changes for PCK-means and MPCK-means are not obvious. Interestingly, even with the number of labeled data growing, Semi-LP still performs worse than Un-LP. As is shown in Figure 3, Un-LP benefits much from multi-exemplars ($e \geq 2$). For D4, Un-LP is capable of achieving 99.58% in F-value when $e = 7$, obtaining 21.32% improvement over the baseline ($e = 1$). This indicates that our algorithm leverages the additional exemplars effectively.

4.5 Case Study

We conduct an experiment to illustrate the characteristics of the proposed algorithm in this subsection. We cluster the words in all the four domain datasets and select the most representative words for each cluster by sorting y_i. In the experiment, we set $L = 4$ in order to match the class number of the dataset. Table 1 shows top-20 representative words for each cluster, where the bold words are the ones

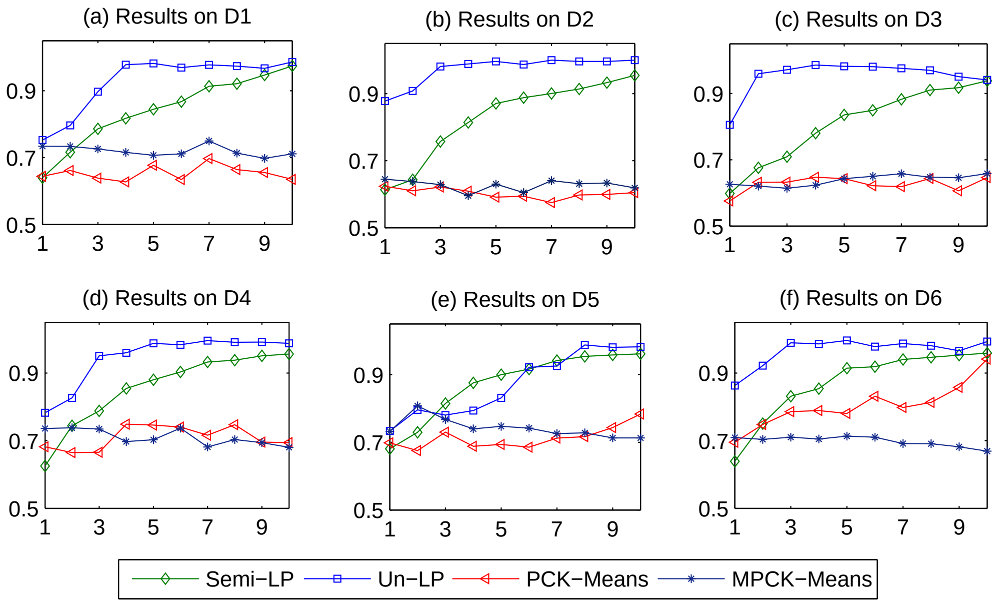

Figure 3: Results on 20NG where X-axis is e, Y-axis is F-value.

domain	*meng*	*configure*	*johansson*	*aug*	geode	isbn	bug	parity
Atheism	100.00%	0	0	0	0	91.67%	89.47%	0
Hardware	0	90.91%	0	0	0	0	10.53%	66.67%
Hockey	0	9.09%	100.00%	0	0	8.33%	0	0
Space	0	0	0	100.00%	100.00%	0	0	33.33%

Table 2: Distributions of the incorrect words partitioned by the literal meaning.

with correct cluster label inferencing from the literal meaning. We observe that the accuracy of word clustering on 20NG is very low (28.75%), which is at variance with the preceding conclusion. One reason is that words in 20NG are stemmed, so, from Table 1 it can be clearly seen that there are some non-English words (e.g., ”mcl”, ”hfd”, ”stl”, etc.) that don’t have actual meanings.

In order to gain further insights into the reasons, the distributions of these incorrect words have been made in statistics. Partial results are shown in Table 2. From the distributions, we can find that many words marked in italics in Table 1 have been correctly clustered, although they have nothing to do with corresponding class in the literal meaning. Taking these words into account, the accuracy can reach 81.25% which demonstrates once again the effectiveness of Un-LP word clustering algorithm.

5 Conclusion

In this paper, we propose an unsupervised label propagation algorithm to tackle the problem of word clustering. The proposed algorithm uses a similarity graph based on co-occurrence information to encourage similar words to have similar cluster labels. One of the advantages of this algorithm is that it uses multi-exemplars to represent a cluster, which can significantly improve the clustering results.

Acknowledgements

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (XDA06030602), National Nature Science Foundation of China (No. 61202226), National 863 Program (No. 2011AA010703), IIE Program (No.Y3Z0062201).

References

Basu S, Bilenko M, Mooney R J. 2004. A probabilistic framework for semi-supervised clustering. In *Proceedings of SIGKDD*, pages 59-68.

Bilenko M, Basu S, Mooney R J. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of ICML*.

Blei D M, Ng A Y, Jordan M I. 2003. Latent dirichlet allocation. *The Journal of machine Learning research*, pages 993-1022.

Candito M, Seddah D. 2010. Parsing word clusters. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 76-84.

Han H, Manavoglu E, Zha H, et al. 2005. Rule-based word clustering for document metadata extraction. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1049-1053.

Jeff M A, Matsoukas S, Schwartz R. 2011. Improving Low-Resource Statistical Machine Translation with a Novel Semantic Word Clustering Algorithm. In *Proceedings of the MT Summit XIII*.

Jin P, Sun X, Wu Y, et al. 2007. Word clustering for collocation-based word sense disambiguation. *Computational Linguistics and Intelligent Text Processing*, pages 267-274.

Koo T, Carreras X, Collins M. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-HLT*, pages 595-603.

Krause A, Gomes R G. 2010. Budgeted nonparametric learning from data streams. In *Proceedings of ICML*, pages 391-398.

Momtazi S, Klakow D. 2009. A word clustering approach for language model-based sentence retrieval in question answering systems. In *Proceedings of CIKM*, pages 1911-1914.

Sagae K, Gordon A S. 2009. Clustering words by syntactic similarity improves dependency parsing of predicate-argument structures. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 192-201.

Sun A, Grishman R, Sekine S. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of ACL*, pages 521-529.

Turian J, Ratinov L, Bengio Y. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL*, pages 384-394.

Uszkoreit J, Brants T. 2008. Distributed Word Clustering for Large Scale Class-Based Language Modeling in Machine Translation. In *Proceedings of ACL*, pages 755-762.

Wu Q, Ye Y, Ng M, et al. 2010. Exploiting word cluster information for unsupervised feature selection *Trends in Artificial Intelligence*, pages 292-303.

Zhu X, Ghahramani Z. 2002. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University*.

Zhu X, Ghahramani Z, Lafferty J. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of ICML*, pages 912-919.