The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Adaptation of Multivariate Concept to Multi-Way Agglomerative Clustering for Hierarchical Aspect Aggregation

Tamasha Malepathirana,[1] Rashindrie Perera,[1] Yasasi Abeysinghe,[1]
Yumna Albar,[1] Uthayasanker Thayasivam[1]
[1]Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
tamasha@cse.mrt.ac.lk, rashindrie.14@cse.mrt.ac.lk, yasasi.14@cse.mrt.ac.lk.,
yumna.14@cse.mrt.ac.lk, rtuthaya@cse.mrt.ac.lk

Abstract

Hierarchical review aspect aggregation is an important challenge in review summarization. Currently, agglomerative clustering is widely used for hierarchical aspect aggregation. We identify an important but less studied issue in using agglomerative clustering for the aforementioned task. This paper proposes a novel approach to generate a multi-way hierarchy by adaptation of the multivariate concept. Furthermore, we propose a novel experimentation approach to evaluate the acceptability of the aspect relations obtained from the hierarchy generated.

Introduction

With the large amount of user reviews available online, content grasping has become a major concern in today's world. Obtaining a summary of reviews of a product or service has therefore become a research interest in many research works (Yu et al. 2011; Anand et al. 2018). Reviews generally contain aspects which are the features of the target upon which consumers comment. Identifying the latent structure among aspects would benefit the process of summarizing reviews. Even though aspect based review summarization is a thoroughly researched topic, most of the research so far ignores the hierarchical structure of the aspects which is in fact of great importance (Kim et al. 2013). Some research work done in terms of aspect hierarchy generation follow supervised or semi-supervised approaches (Carenini, Ng, and Zwart 2005; Yu et al. 2011) requiring domain assisted knowledge or unsupervised approaches (Pavlopoulos and Androutsopoulos 2014; He et al. 2015) which mostly use agglomerative clustering.

The main drawback of using agglomerative clustering is the resultant binary tree which fails to cater the requirements of almost all applications. Binary tree introduces a range of limitations to the process. Firstly, as the number of aspects increases, the height of the tree increases considerably. The taller the tree, the larger the number of steps the user has to perform to reach to a desired aspect. Secondly, the inability to represent more than two aspects at the same level. This limitation results in loss of information as it fails to represent relationships where each aspect could have more than 2 sub-aspects. Thirdly, the resultant hierarchy is not a regular form of hierarchy that a user would prefer to see.

Moreover much research work has been carried out to study the use of word embedding to the task of aspect aggregation and thus have obtained promising results as opposed to traditional approaches (Xiong and Ji 2016; Ye et al. 2015). However, the resulting vectors are of high dimensions which questions the applicability of currently proposed clustering solutions. Our focus on this paper is to present a novel approach to address above issues by modifying the traditional agglomerative clustering algorithm incorporating the concept of multi-variate generalization of standard deviation. We introduce the concept of Mahalanobis distance for the process of hierarchical aspect aggregation. We also introduce a subjective evaluation method for the resultant hierarchy. Even though our approach can be used for automatically building a concept hierarchy in any domain, in this paper our focus has been on the restaurant domain.

Related Work

Traditional agglomerative clustering has been used for hierarchical aspect aggregation (Pavlopoulos and Androutsopoulos 2014). However, the above research work produces a binary tree facing the limitations mentioned in Introduction . Huang et al. (Kuo and Huang 2005) who identified the limitation of traditional agglomerative clustering, proposed a multiple-way agglomerative hierarchical clustering algorithm by introducing a new join operation instead of the default merge operation. The new join operation allowed a cluster to join a cluster at the upper level such that the cluster of the upper level is the parent of the cluster of the current level. Kuo et al. (Kuo, Tsai, and Huang 2006) later extended the above research by further modifying the traditional agglomerative clustering method. In their work, once the two closest clusters are selected for combining into a new cluster, algorithm decides whether to create a new cluster with the two original clusters as its sub-clusters, or to perform a join operation by merging their children i.e., sub-clusters, into a single cluster. Tu et al. (Tu, Chen, and Chen 2015) proposed an introduction of a partitioning process based on the KNN connected graph to overcome the drawback of not obtaining

INITIALIZATION:
C = set of clusters per each $a_i \epsilon A$
D = set of link distances between$a_i, a_j \epsilon A \Lambda a_i \neq a_j$
WHILE C.size >1 :
X = min(D).left_cluster Y = min(D).right_cluster
Create cluster C_n&add to C
IF (ACD(X,Y)-WACD(X,Y)> α OR X AND Y leaves:
$addX\&Yassubclustersof C_n$
ELSE IF ($\|Dist(X) - Dist(Y)\|/WACD(X,Y) < \beta$:
add subclusters of $X\&Y$ as subclusters of C_n
remove $X\&Y$ from C
modified_cluster = C_n
ELSE IF (Dist(X)>Dist(Y)) :
Add X as a subcluster of Y
Replace C_n by Y & remove C_n from C
modified_cluster = Y
ELSE :
Add Y as a subcluster of X
Replace C_n by X & remove C_n from C
modified_cluster = X
Adjust Dist(modified_cluster)
Modify_link_distances(D, modified_cluster)
END WHILE

Table 1: Proposed agglomerative clustering algorithm

the optimum in the global space.

Methodology

In this section, we present a detailed description of our algorithm proposed to modify the traditional agglomerative clustering algorithm.

Input

The input to the proposed algorithm is a set of aspect vectors which are obtained by amalgamating the pre-trained Google News corpus Word2Vec model and domain specific word embedding models trained using combined review corpora.

Algorithm

In traditional agglomerative clustering algorithm each observation is assigned to its own cluster. Then, between each cluster pair, a similarity is computed and the two most similar clusters are merged. In order to calculate the similarity, several linkage methods are used in the literature including Single, Complete, Ward and Average linkage. The process is repeated until a single cluster is formed.

The proposed algorithm is shown in Table 1. In contrasting to the traditional agglomerative clustering algorithm, we use two types of join operations proposed by Huang et al.(Kuo, Tsai, and Huang 2006) in addition to the merge operation. The types of operations used in the algorithm are explained below.

1. Merge Operation: Default operation used in traditional agglomerative clustering.
2. Level Up Join: Sub-clusters of cluster X and subclusters of cluster Y are joined to be the sub clusters of a new cluster.
3. Level Down Join: One cluster becomes a sub cluster of another.

Operations 2 and 3 above, allows a cluster to have more than 2 subclusters. We first introduce the below definitions. Let X and Y be two clusters consisting *m* and *n* objects respectively . Assume $\{X_1, X_2, ..., X_p\}$ are sub clusters of X and $\{Y_1, Y_2, ..., Y_q\}$ are subclusters of Y.

Definition 1 *ACD(X,Y) : Averaged Cluster Distance of X and Y, is the average distance between an object in X and an object in Y.*

$$ACD(X,Y) = \frac{\sum_{O_i \in X}^{m} \sum_{O_j \in Y}^{n} Dist(O_i, O_j)}{m * n} \quad (1)$$

Definition 2 *TCD(X) - Total Cluster Distance of X, is the total distance between subcluster pairs in X.*

$$TCD(X) = \sum_{X_i \in X}^{p} \sum_{X_j \in X \wedge X_i \neq X_j}^{p} ACD(X_i, X_j) \quad (2)$$

Definition 3 *ATCD(X) - Averaged Total Cluster Distance of X, is the TCD(X) divided by the number of subcluster pairs.*

$$ATCD(X) = \frac{\sum_{X_i \in X}^{p} \sum_{X_j \in X \wedge X_i \neq X_j}^{p} ACD(X_i, X_j)}{\frac{p(p-1)}{2}} \quad (3)$$

Definition 4 *WACD(X, Y) - Weighted Averaged Cluster Distance between X and Y.*

$$WACD(X,Y) = \frac{TCD(X) + TCD(Y)}{\frac{p(p-1)}{2} + \frac{q(q-1)}{2}} \quad (4)$$

Determination of the Adjusted Cluster Distance of X (*AdCD(X)*)

The graphical representation of the agglomerative clustering output is a dendrogram where the node height represents the linkage distance between the two sub clusters. But in the modified clustering algorithm, a node is entitled to have more than two children failing to obtain the cluster height from the traditional way. Thus, a new mechanism is needed to obtain the new cluster height after performing a Level up or a Level down join operation.

Mahalanobis distance(Mahalanobis 1936) measures the number of standard deviations away a point is from the mean of its distribution. Mahalanobis distance is used in our work to incorporate multi-dimensional generalization of the idea in measuring the distance between a cluster to its distribution,

1. In determining the height/distance of the modified cluster
2. In determining the best possible threshold to perform the join operations

Mahalanobis Distance accounts for the aspect characteristics such as high dimensionality, variance difference among vector dimensions and co-variance among aspect dimensions. Thus, introducing the concept of multivariate standard deviation to the task.

We take the new height of the modified cluster to be the weighted average of the sub-cluster heights and its own

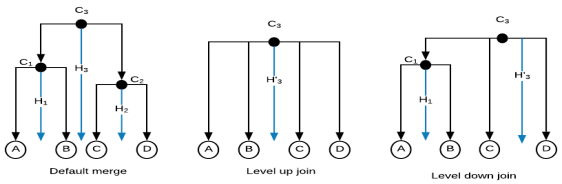

Figure 1: Join operations for clusters C_1 and C_2: Default merge, level up and level down join operations

height. The heights are weighted by the reciprocal of the Mahalanobis distance from a cluster to its distribution which is defined as the distance from the average of cluster leaves (aspect vectors) to the distribution (D). Thus, the weighting factor W can be calculated as shown in Equation 5.

$$W(X) = \frac{1}{MD(\frac{\sum_{O_i \in X} O_i}{m}, D)} \tag{5}$$

The adjusted cluster distance can then be obtained as follows in Equation 6.

$$\frac{\sum_{X_i \in X} AdCD(X_i) * W(X_i) + AdCD(X) * W(X)}{Normalization factor} \tag{6}$$

The new height of C_3 is determined by the weighted average of clusters C_1 , C_2 and C_3. Intuition behind weighting the heights by the reciprocal of Mahalanobis distance is to capture the goodness of existing clusters in to the new height. If the Mahalanobis distance from C_1 to the distribution is lower that means the average distance of leaves in C_1 is closer to the mean of the distribution. Thus, a higher portion of C_3's height corresponds to the height of C_1. Similarly, if the Mahalanobis distance from C_2 to the distribution is higher that means the average distance of leaves in C_2 is further away from the mean of the distribution. Thus, only a lower portion of C_3's height corresponds to the height of C_2.

Determination of threshold values

In the original algorithm, threshold is taken to be the standard deviation. But for multi-dimensional data, the representation of standard deviation from a single value is futile. Therefore, we introduce a new cost function to determine the threshold value for which the tree structure returns the lowest cost. The cost of a cluster is calculated by taking the Mahalanobis distance from the cluster mean vector to the distribution divided by the number of aspects in that cluster.

$$Cost(X_i) = \frac{MD(\frac{\sum_{a_i \in X} a_i}{m}, D)}{m} \tag{7}$$

The cost of the whole tree structure is calculated iteratively taking the sum of costs in all the clusters formed. Starting from the root, we recursively call the cost function for all resulting subclusters. Threshold values are taken to be the values for which the *Cost(root)* is minimum.

	Annotator 1	Annotator 2	Annotator 3	Annotator 4
Traditional Algorithm	0.2656	0.1732	0.1928	0.2773
Modified Algorithm	0.6632	0.4179	0.4968	0.4589
Improvement	3.8x	2.4x	2.6x	1.65x

Table 2: Spearman's Rank Coefficient Correlation values obtained for traditional and proposed algorithm

Experiments

This section contains the experiments carried out to evaluate the proposed algorithm. We aim to measure the acceptability of the resulting hierarchy as opposed to the traditional agglomerative clustering hierarchy for the task of aspect aggregation. The experiment was done using a set of aspects provided by Pavlopoulos and Androutsopoulos (Pavlopoulos and Androutsopoulos 2014). In order to exhibit the hierarchical structure among aspects more vividly, we added few additional aspects to the initial dataset.

Due to the unavailability of a true hierarchy and the subjectiveness of hierarchies designed by humans, we introduce a novel experiment approach to measure the exactitude of the resulting hierarchy. Firstly, for the same set of aspects we obtained 4 hierarchies designed by 4 different human annotators. Then each hierarchy is compared with the outputs of the traditional and modified algorithms. To compare two hierarchical structures we introduce a novel approach using the concept of Lowest Common Ancestor (LCA) . To compare two hierarchies H_1 and H_2 we build an upper triangular matrix A for each hierarchy where A[i,j] represent the LCA distance between the aspect in row i and column j. In literature, the LCA of two leaves X and Y is defined to be the node of greatest depth that is also an ancestor of both X and Y (Ganesan, Garcia-Molina, and Widom 2003). LCA distance between two leaves X and Y is computed by taking the sum of each aspect's distance to their LCA. We then measure the similarity between the values in two matrices using Spearman's Rank Coefficient Correlation.

Results and Analysis

As shown in the Table 2, for each test case the proposed algorithm has obtained higher values for the similarity with the expected values than the traditional agglomerative clustering algorithm. Since the proposed algorithm allows a cluster to have more than or equal to two subclusters, it overcomes the limitation of traditional agglomerative clustering algorithm being binary. The proposed algorithm is capable of capturing the distance between high dimensional vectors more effectively generating more meaningful clusters. The variation between correlation values among annotators can be due to the subjectiveness of the hierarchy. Therefore, we also looked into the annotator agreement between each annotator pair using Spearman's Rank Correlation Coefficient and the resulted values which were varying in the range of 0.4-0.8 show the subjectiveness of the hierarchy. That is, we cannot strictly identify one hierarchical structure to be

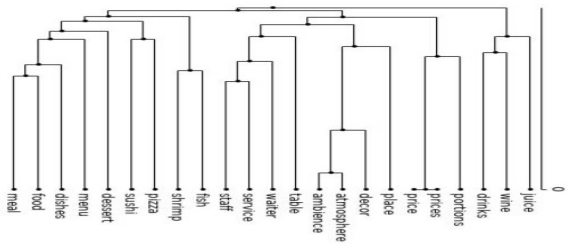

Figure 2: Dendrogram generated by traditional algorithm

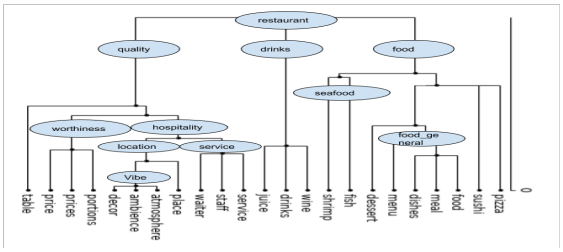

Figure 3: Dendrogram generated by proposed which is a multi-way hierarchy as opposed to the binary hierarchy generated for the same set of aspects in Figure 2.

correct and the other to be wrong. Each hierarchy depends on individual's perspective. Thus, what should be evaluated is the acceptability of the aspect relations obtained. Getting a substantial similarity score between the annotated hierarchies and the obtained hierarchy helps to achieve the objective.

Conclusion

We address an important but not well studied issue in using agglomerative clustering for hierarchical aspect aggregation. We identify the importance of a multi-way hierarchical structure in contrast to that of a binary structure for the task of multi granualr aspect aggregation. We show that the modification of the agglomerative clustering algorithm by adapting the Mahalanobis distance can help cluster aspects to produce more meaningful and user desirable multi granular aspect hierarchies. Finally, we propose an evaluation method to measure the acceptability of the aspect relations depicted in the obtained hierarchy by utilizing user desired hierarchies.

References

Anand, K.; Dewangan, N.; Kumar, N.; and Singh, M. 2018. Aspect ontology based review exploration. *Electronic Commerce Research and Applications*.

Carenini, G.; Ng, R. T.; and Zwart, E. 2005. Extracting knowledge from evaluative text. In *Proceedings of the 3rd International Conference on Knowledge Capture*, K-CAP '05, 11–18. New York, NY, USA: ACM.

Ganesan, P.; Garcia-Molina, H.; and Widom, J. 2003. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)* 21(1):64–93.

He, Y.; Song, J.; Nan, Y.; and Fu, G. 2015. Clustering chinese product features with multilevel similarity. In Sun, M.; Liu, Z.; Zhang, M.; and Liu, Y., eds., *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 347–355. Cham: Springer International Publishing.

Kim, S.; Zhang, J.; Chen, Z.; Oh, A. H.; and Liu, S. 2013. A hierarchical aspect-sentiment model for online reviews. In *AAAI*.

Kuo, H.-C., and Huang, J.-P. 2005. Building a concept hierarchy from a distance matrix. In *Intelligent information processing and web mining*. Springer. 87–95.

Kuo, H.-C.; Tsai, T.-H.; and Huang, J.-P. 2006. Building a concept hierarchy by hierarchical clustering with join/merge decision. In *JCIS*.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. National Institute of Science of India.

Pavlopoulos, J., and Androutsopoulos, I. 2014. Multigranular aspect aggregation in aspect-based sentiment analysis. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 78–87. ACL.

Tu, D.; Chen, L.; and Chen, G. 2015. Automatic multiway domain concept hierarchy construction from customer reviews. *Neurocomputing* 147:472–484.

Xiong, S., and Ji, D. 2016. Exploiting flexible-constrained k-means clustering with word embedding for aspect-phrase grouping. *Information Sciences* 367-368:689 – 699.

Ye, K.; Li, L.; Guo, M.; Qian, Y.; and Yuan, H. 2015. Summarizing product aspects from massive online review with word representation. In Zhang, S.; Wirsing, M.; and Zhang, Z., eds., *Knowledge Science, Engineering and Management*, 318–323. Cham: Springer International Publishing.

Yu, J.; Zha, Z.-J.; Wang, M.; Wang, K.; and Chua, T.-S. 2011. Domain-assisted product aspect hierarchy generation: Towards hierarchical organization of unstructured consumer reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 140–150. Stroudsburg, PA, USA: Association for Computational Linguistics.