Ontology Construction for Eldercare Services with an Agglomerative Hierarchical Clustering Method

Peng Han[1], Yulong Li[1], Yue Yin[2(✉)], and Ning An[1]

[1] School of Computer and Information, Hefei University of Technology, Hefei, China
[2] Institute of Industrial and Equipment Technology, Hefei University of Technology, Hefei, China
frank.y.yin@qq.com

Abstract. A high-quality ontology for eldercare service can help deliver high quality eldercare services in increasingly aged and digitized societies because it can serve as a reference for formulating eldercare service standards and exchanging information pertain to eldercare services over the Internet. Improving upon previous work, we proposed an agglomerative hierarchical clustering method to construct such an ontology. This method incorporates longitudinal denoising and the word bag model to achieve accurate results verified by experiment results.

Keywords: Ontology construction · Eldercare services · Agglomerative hierarchical clustering

1 Introduction

As people live in increasingly aged and digitized societies, many researchers have looked into assisting elders to cope with the digital world [1, 2]. However, little research has studied how to provide standard names for eldercare services that have been growing in both volume and varieties. Many eldercare service names are by convention or coined by service providers. This leads to at least two issues: (1) one service has multiple names and (2) one name means different services at different places. These issues could cause confusion in the marketplace and hinder the information exchange pertain to eldercare services on the Internet.

As shown in other fields [3], an effective way to tackle this problem is to build an ontology for eldercare services. Ontology is the philosophical study of being. More broadly, it studies concepts that directly relate to being, in particular becoming, existence, reality, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology often deals with questions concerning what entities exist or may be said to exist and how such entities may be grouped, related within a hierarchy, and subdivided according to similarities and differences. Philosophers can classify ontologies in various ways, using criteria such as the degree of abstraction and field of application [4]:

J. Zhou and G. Salvendy (Eds.): HCII 2019, LNCS 11592, pp. 34–45, 2019.
https://doi.org/10.1007/978-3-030-22012-9_3

Upper ontology, Domain ontology, Interface ontology and Process ontology. Here we choose the domain ontology, for example, information technology, computer languages and the specific branches of science. The ontology defines a common dictionary for researchers to use when sharing information in a particular domain. It contains the basic concepts in the field that can be interpreted by the machine and the relationships between the concepts. Domain experts can leverage many rules for developing standardized ontology. Building an ontology is like defining a set of data and their structure for use by other programs. At present, there are several cases in which other ontology libraries have been successfully constructed in other fields, for example, Ontology Construction for Information Selection [5], Ontology construction for information classification [6] and so on. The use of these tools significantly reduces the complexity of dealing with patients at home [7]. To our best knowledge, we were the first to propose building an ontology for eldercare services in China [8]. The K-means method was used to preliminary studied building an ontology for eldercare service in China, but there is no evaluation of the clustering result and the related names merging. In the previous method, they used the K-means method which used must give an initial cluster number K first, the result of clustering will be different for different K values. When there is noise in the dataset, the k-means clustering proceed will deviate considerably. In this paper, we perform the text denoising before we apply the agglomerative hierarchical clustering method.

The construction of ontology is a systematic project that should follow certain construction guidelines and under the guidance of reasonable methodology, using suitable ontology description language and convenient ontology development tools [9]. There are many methods to build an ontology, for example, Skeletal Methodology, IDEF5 (Integrated Definition for Ontology Description Capture Method), TOVE, Methontology, and seven-step method. In this paper, we use the seven-step method. The seven-step method was developed by Stanford University School of Medicine and is mainly used for the construction of domain ontology, they are determined the scope and scope covered by the ontology, considering multiplexing existing ontology, enumerate all terms, defining the hierarchy of a class, define the properties of the class, define class constraints, create instances. In the seven-step method of ontology construction, the most critical step is the fourth step. There are usually many concepts involved in building ontology, and manually defined methods are often inefficient. Everyone has a different understanding of a particular term may result in multiple classifications. In view of this, many researchers will do preliminary classification through cluster analysis.

2 Clustering Methods

Clustering is to divide a given data set into different clusters according to a specific criterion, so that the similarity between objects in the same cluster is high whereas the dissimilarity between objects in the different clusters is high [10].

Researchers generally categorize clustering algorithms into the following six categories: partition-based methods, density-based methods, network-based methods, model-based methods, fuzzy-based clustering and hierarchy methods.

Partition-based methods [11]: firstly, it is necessary to determine the number of clusters, then select several points as the initial center point randomly, and iteratively reset the points according to the predetermined heuristic algorithm until the target effect is finally achieved. Partition-based methods are simple and efficient for large data sets, furthermore, it has good performance both in time and space. The disadvantage is that the result is easy to be locally optimal when the data set is large, the K value needs to be set in advance and it is very sensitive to the K points selected at the beginning. Such algorithms form the k-means algorithm and its variants, including k-medoids, k-modes, k-medians, kernel k-means, and other algorithms.

Density-based methods [12]: there are two parameters should be defined, one is the maximum radius of the circle, the other is the number of points that the circle should contain at least. As long as the density of adjacent regions (the number of objects or data points) exceeds a certain threshold, the clustering will continue. Finally, each circle corresponds a class. Its advantage is that it is insensitive to noise and can find the arbitrary shapes. The disadvantage is that the result of clustering has a great relationship with parameters. Its typical algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

Grid-based methods [13]: the principle of this method is to divide the data space into grid cells, map the data object set into grid cells, and calculate the density of each cell. According to the preset threshold value, each grid cell is judged to be a high-density cell, and a class is formed from a group of adjacent dense cells. The advantage of this type of method is that it is fast because its speed is independent of the number of data objects but only depends on the number of cells in each dimension in the data space. The disadvantages are sensitive to parameters, inability to handle irregularly distributed data.

Model-based methods [14]: this method assumes a model for each cluster and looks for the best fitting of data from the given models. This kind of method mainly refers to probabilistic model-based method and neural network model-based method, especially probabilistic model-based method. Its advantage is that the division of classes is expressed in the form of probability, and the characteristics of each class can also be expressed by parameters. The disadvantage is that the execution efficiency is not high, especially when the distribution quantity is large and the data quantity is small. The most typical and commonly used method is the GMM (Gaussian Mixture Models).

Fuzzy based clustering [15]: a sample belongs to a class with a certain probability. There are some typical fuzzy clustering methods based on objective function, similarity relation and fuzzy relation. The FCM algorithm is an algorithm that determines the degree to which each data point belongs to a certain cluster degree by membership degree. This clustering algorithm is an improvement of the traditional hard clustering algorithm. It is sensitive to isolated points, but it does not ensure that the FCM converges to an optimal solution.

Hierarchical methods are described in more detail in Sect. 3.

Because this paper study the name of eldercare services, we focus on text clustering here. Text clustering is to treat every text as a sample and clustering all the samples. However, the difference from the common clustering operations in machine learning is that the text clustering object is not the direct text itself, but the features extracted from the text. Because the characteristics of the text directly affect the quality of the

clustering results, so how to extract features and filter noise is a very important step [16]. There are two kinds of algorithms for text clustering. The first kind is based on layered algorithms, including single links, full links, group averaging, and Ward methods. Using aggregation or division method, one can cluster documents into a hierarchical structure. The other kind contains K-means and its variants. In general, layered algorithms produce more in-depth information for detailed analysis, while algorithms based on variants of the K-means algorithm are more efficient and provide sufficient information for most purposes. Here we build the corresponding text documents by merging the search results of the care service names, and the clustering result of the corresponding service name is obtained by clustering the text documents.

3 Proposed Hierarchical Clustering Method

In this paper, we apply an agglomerative hierarchical clustering method to construct an ontology for eldercare services. We first acquire names of the eldercare service by assigning search engine results as a text document for the corresponding service name. Next, we segment these text documents before denoising and longitudinal denoising them. Then, we use the TF-IDF method to give weights of all word segmentation results of the corresponding name. Subsequently, converts each eldercare service name into a word vector. At last, we apply an agglomerative hierarchical clustering method to cluster the service names in multiple layers. Figure 1 illustrates the follow of the proposed method.

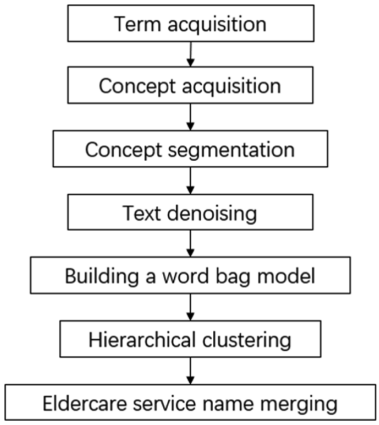

Fig. 1. Eldercare service name clustering method flow

3.1 Term Acquisition

The eldercare service names under study come from two sources. The first group of eldercare services was crawled from the website of the civil affairs department of each province and city. This group of service names include meal service, living service,

bathing service, sanitation cleaning service, agency service, preventive health service, medical assistance service, health consultation service, elderly health file filing. The second group of service names came from the official websites of major pension institutions, nursing homes and pension service companies including health assessment services, physical examination services, health knowledge courses, self-health management services, food supplements, cleaning, walking, bathing, lighting, hood cleaning, carpet cleaning, air conditioning cleaning, errands, vegetables and fruits. The total number of names is 412. It is clear that many service names could overlap with others in terms of services they represent, such as babysitters and home sitters, cleaning and sanitation.

3.2 Concept Acquisition

In the Baidu search engine, we crawl the search results for each eldercare service name. The search results include a subtitle of each column and an introduction below the title. We use 20 search results for each service name and merge them together as a text document corresponding to the service name. Because the content of this advertisement may be more relevant to the service, the search results contain the advertisements.

3.3 Concept Segmentation

We use the jieba tokenizer to perform word segmentation on our text dataset. The jieba algorithm for word segmentation uses a prefix-based dictionary to achieve efficient word graph scanning, and generates a directed acyclic graph composed of all possible word generated by Chinese characters in sentences. Dynamic programming is used to find the maximum probability path, and the maximum segmentation combination based on word frequency is found. For the unregistered words, the HMM model based on the ability of Chinese characters is adopted. The jieba tokenizer is used because it is suitable for search engine segmentation. Using the precise model, the long words are divided into words and the recall rate is improved.

The jieba word segmentation actually supports custom word segmentation, so we need to add our care service names to the custom dictionary in advance to avoid splitting our care service names. Finally, more than 1000 kinds of Chinese stop words in the network are used to perform preliminary noise reduction on the processed word segmentation text documents.

3.4 Text Denoising

Through the pre-processing of the previous steps, we have obtained a relatively complete text: the first column of each line is the name of the eldercare service and remaining columns are the search results related to this particular service name. Longitudinal denoising is achieved by comparing the similarity between the service name and each column of word segmentation, then intercept the similarity higher than the set threshold. The word2vec model proposed by Mikolov et al. [17] has attracted a great amount of attention in recent years, and it has shown effectiveness in text mining analysis. In fact, it is a group of related models used to produce word embedding.

Taking a large corpus of text as input, the Word2vec generates a vector space that typically is of several hundred dimensions, before it positions each unique word in the vector space in such way that words that share common contexts in the corpus are located in close proximity to one another in the space [17]. Through the training of a large number of corpora, each word is mapped into a vector of high dimension, that is, the processing of text content is simplified to vector operation in vector space, and the similarity in vector space is calculated to indicate the similarity between words. By setting the similarity threshold, one can delete many unrelated words to improve the clustering effect. In this paper, the result segments with thresholds of 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, and 0.9 are set as new texts, the best clustering results are obtained by comparison.

3.5 Building a Word Bag Model

The next process needs to perform the following steps on the seven data sets obtained in the previous step and compare them for the best results.

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as the bag of its words (disregarding grammar and even word order but keeping multiplicity). The bag-of-words model is commonly used in methods of document classification where the occurrence of each word is used as a feature for training a classifier. Zellig Harris's 1954 paper on Distributional Structure [18] is an early reference to "bag of words" in a linguistic context. Intuitively, all the word segmentation results in this paper constitute a large word bag, each eldercare services have a corresponding word bag with different value.

The term "Term Frequency–Inverse Document Frequency" (abbreviated as TF-IDF) is a commonly used weighting technique for information retrieval and data mining and it intends to reflect how important a word is to a document in a collection or corpus [19]. The importance of words increases with the number frequency appear in the file, but decreases inversely as it appears in the corpus. Various forms of TF-IDF weighting are extensively applied by search engines as a measure or rating of the degree of correlation between a file and a user query.

If a word or phrase appears in an article with a high frequency and rarely appears in other articles, then the word or phrase is considered to have good class distinguishing ability and is suitable for classification. For the word t_i in a particular file, its weight can be calculated as follows:

The word frequency (TF) refers to the frequency at which a given word appears in the file. The value is obtained by the following formula:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Where: $n_{i,j}$ denotes the number of occurrences of the word in the file d_j.

$\sum_k n_{k,j}$ denotes the sum of the occurrences of all the words in the file d_j.

The inverse document frequency (IDF) is a measure of the universal importance of a word. The value is obtained by the following formula:

$$idf_i = \log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right|} \tag{2}$$

Where: $|D|$ denotes the total number of files in the corpus

$\left|\left\{j : t_i \in d_j\right\}\right|$ denotes number of files containing the word t_i (the number of files with $n_{i,j} \neq 0$) will cause the dividend to be zero if the word is not in the corpus.

Then calculate the product of TF and IDF:

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

The high word frequency within a particular file, and the low file frequency of the word in the entire file set, can produce a high weight of TF-IDF. Therefore, TF-IDF tends to filter out common words to retain important words.

The result is that each service name corresponds to a word vector, and each vector has the same dimension.

3.6 Hierarchical Clustering

It groups a set of data in a way that maximizes the similarity within clusters and minimizes the similarity between two different clusters. Hierarchical clustering methods seek to build a hierarchy of clusters. The strategies for hierarchical clustering generally contain two types:

Divisive: a top-down approach starts from all the texts, and splits the texts according to some rules recursively.

Agglomerative: a bottom-up approach considers each text as a cluster, and merges pair of clusters for each step.

We chose the agglomerative method in this paper. Its basic ideas are as follows. Initially each sample is treated as a cluster, so the size of the original cluster is equal to the number of samples, and then these initial clusters are merged according to certain criteria until a certain condition is reached or the set number of categories is reached. The main steps are as follows:

(1) Get the distance matrix of all samples;
(2) Treat each sample as a separate cluster;
(3) Calculate the cosine similarity between each pair of samples (x,y) according to the following formula, and find a pair of sample points with the smallest cosine similarity (x,y):

$$cos(x, y) = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} * \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{4}$$

Where: x, y denote any two clusters, also refer to the service name;

x_i denotes the weight corresponding to each word in the word bag after the x cluster;

(4) Combine the two clusters x and y with the largest cosine similarity, then update the distance matrix of the merged samples by calculating the mean of these two matrixes;
(5) Repeat steps (2)–(4) until we merge all samples into one cluster.

The advantage of hierarchical clustering methods is that it does not require specifying the number of clusters in advance. On the contrary, K-means method used in our previous study [8] must first give an initial cluster number K, and the result of clustering will depend on different K values. Hierarchical clustering methods view data having different granularity levels, hence can assist people to visualize and interactively explore large document collections [20], then we cut the clustering tree according to the quality of clusters.

However, the disadvantage of hierarchical clustering is that the amount of computation is large, since hierarchical clustering algorithms require pairwise inter object proximities, the complexity of these clustering procedures is at least $O(N^2)$ [21].

3.7 Eldercare Service Name Merging

Researchers intervene at this step and divide the clustering results into 24 named subcategories. The training model introduced in step 3.4 is then used to calculate the similarity of the two care services in each category. The calculation results will provide a reliable reference to decide whether to merge.

4 Experimental Results and Analysis

4.1 Result Verification

According to the clustering results, the final clustering results have 24 categories. In order to select the optimal clustering results, we calculated the F-measure of 9 corresponding result types in 3.4. We select 2 researchers to manually label the 412 care service names according to the naming results, and calculate the precision, recall and F-Measure. The accuracy rate is the proportion of objects that are clustered in the clustering result, and it is judged whether each service name is correctly classified to the category to which it belongs. The recall rate is the proportion of similar names in the same topic merged into a class, and it is judged whether the name of each category is completely divided into the corresponding class.

F-measure is a composite indicator that is the harmonic average of precision and recall. The larger the value of f-measure is, the better the clustering effect is. The formula is as follows:

$$\text{Precision} = \frac{n_{ij}}{n_i} \tag{5}$$

$$\text{Recall} = \frac{n_{ij}}{n_j} \tag{6}$$

$$\mathrm{F-Measure} = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{7}$$

where: n_{ij} denotes the number of names in class i manually labeled in cluster j;

n_i denotes the number of names in manually labeled class I;

n_j denotes the number of names in cluster j.

Finally, the clustering result obtained when the threshold is set to 0.75 is the best. The F-measure results are shown in Table 1. For the sake of simplicity, we have chosen 1–24 instead of 24 types of services.

Table 1. F-measure

Category code	1	2	3	4	5	6
Researcher 1	0.86	0.96	0.86	0.85	0.81	0.9
Researcher 2	0.9	0.92	0.88	0.8	0.86	0.93
Category code	7	8	9	10	11	12
Researcher 1	0.93	0.91	0.96	0.95	0.94	0.9
Researcher 2	0.98	0.93	0.97	0.86	0.88	0.91
Category code	13	14	15	16	17	18
Researcher 1	0.95	0.98	0.95	0.7	0.94	0.86
Researcher 2	0.91	0.96	0.89	0.9	0.97	0.93
Category code	19	20	21	22	23	24
Researcher 1	0.89	1	0.82	0.84	0.8	0.96
Researcher 2	0.95	1	0.94	0.95	0.91	0.82

According to the results, the proposed method consistently achieves the good F-measure. The eldercare service names with obvious characteristics such as hospital departments and health records have higher accuracy.

4.2 Eldercare Service Names Merging Result

Utilizing domain knowledge, we finally merged the eldercare service names: 23 pairs of service names merged and 28 service names deleted. For example, unlocking locks, repairing locks, changing lock cylinders, changing locks, technical unlocking, cost collection, payment of various fees, spiritual comfort, psychological comfort, family doctors, family doctor services, health file establishment, health files, appointments Registration, registration services, hospital registration and other services.

4.3 Result Analysis and Display

Through aforementioned steps, we finally build a bottom-up ontology with protégé4.3 and generate the output in OWL format. The ontology is divided into four layers, the first layer is the eldercare services; the second layer is the four major categories of life demand service, daily demand service, health care service and hospital service. There

are 24 categories in the third floor, which are leisure service, life assistant service, life guidance service for the elderly, life distribution service, errands service, housekeeping service, home cleaning service, payment unlocking service, travel service, home appliance repair service, and elderly rights service, elderly cultural and sports services, elderly care services, health assessment services, health monitoring services, rehabilitation care services, psychological care services, common equipment services, feature testing services, accompanying medical services, health records services, hospital services, elderly data services, hospitals Assistant service. The fourth layer is the name of various eldercare services. Each of these classes and its subclasses are inclusive. Part of the results as shown in Fig. 2.

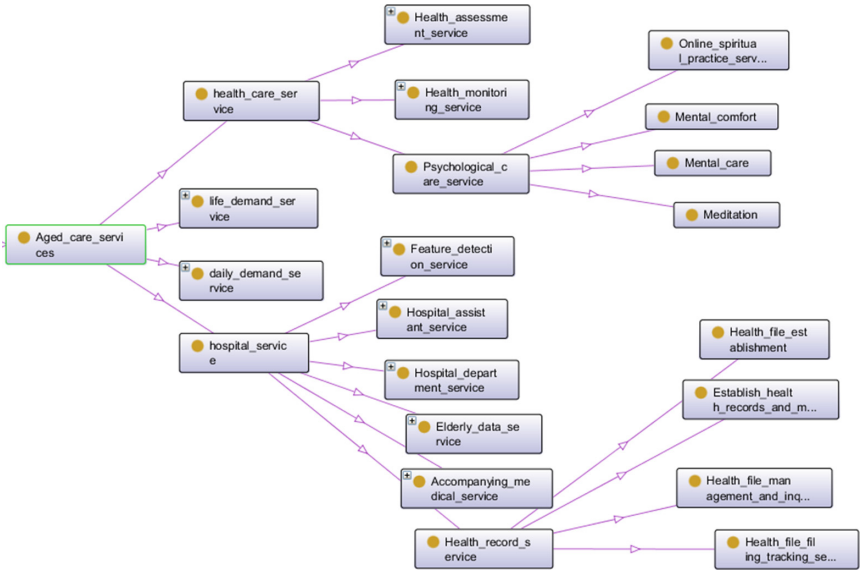

Fig. 2. Diagram of an eldercare service ontology

5 Conclusion

To meet the needs of rapidly aging population, eldercare services continue to grow in both volume and types. While there are new technologies to support this aging population [22], the naming and content of existing, and especially emerging, eldercare services lack of clear specifications. A high quality ontology for eldercare service can help experts categorize these services into standards that not only has their own merits, but also can enable IT technologies to improve the efficiency and quality of eldercare services. Extending our preliminary work on this subject, we propose an agglomerative hierarchical clustering method to improve the quality of the constructed ontology for eldercare services. The experiment results demonstrate the proposed method has a good accuracy. As more work need to be done on this practical and important subject, we

plan further optimize our method, especially by looking into reducing the word bag redundancy.

Acknowledgements. This work was supported in part by the Anhui Key Project of Research and Development Plan under Grant No. 1704e1002221 and the National Program of Introducing Talents of Discipline to Universities ("111 Program") under Grant No. B14025.

References

1. Zhou, J., Rau, P.L.P., Salvendy, G.: Use and design of handheld computers for older adults: a review and appraisal. Int. J. Hum.-Comput. Interact. **28**(12), 799–826 (2012)
2. Wang, A., An, N., Lu, X., Chen, H., Li, C., Levkoff, S.: A classification scheme for analyzing mobile apps used to prevent and manage disease in late life. JMIR mHealth uHealth **2**(1), e6 (2014)
3. Li, X., et al.: Building a practical ontology for emergency response systems. In: 2008 International Conference on Computer Science and Software Engineering, vol. 4, pp. 222–225. IEEE (2008)
4. Petrov, V.: Ontological Landscapes: Recent Thought on Conceptual Interfaces Between Science and Philosophy. Acknowledgements (2011). https://doi.org/10.1515/978311031-9811
5. Khan, L., Luo, F.: Ontology construction for information selection. In: IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society (2002)
6. Weng, S.S., Tsai, H.J., Liu, S.C., et al.: Ontology construction for information classification. Expert Syst. Appl. **31**(1), 1–12 (2006)
7. Riaño, D., Real, F., Campana, F., Ercolani, S., Annicchiarico, R.: An ontology for the care of the elder at home. In: Combi, C., Shahar, Y., Abu-Hanna, A. (eds.) AIME 2009. LNCS (LNAI), vol. 5651, pp. 235–239. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02976-9_33
8. An, N., Yin, Y., Shi, H., Han, P., Cheng, S., Li, L.: Building an ontology for eldercare service in China with a hierarchical clustering method. In: Zhou, J., Salvendy, G. (eds.) ITAP 2018. LNCS, vol. 10927, pp. 3–12. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92037-5_1
9. Li, H., Li, J., Li, M.: Research on modeling method of domain ontology. Comput. Eng. Des. **9**(2), 381–384 (2008)
10. Elhabbash, A.H.: Enhanced k-means Clustering Algorithm (2010)
11. Velmurugan, T., Santhanam, T.: A survey of partition based clustering algorithms in data mining: an experimental approach. Inf. Technol. J. **10**(3), 478–484 (2011)
12. Kriegel, H.P., Kröger, P., Sander, J., et al.: Density-based clustering. Wiley Interdisciplinary Rev.: Data Mining Knowl. Discov. **1**(3), 231–240 (2011)
13. Ma, E.W.M., Chow, T.W.S.: A new shifting grid clustering algorithm. Pattern Recogn. **37**(3), 503–514 (2004)
14. Zhong, S., Ghosh, J.: A unified framework for model-based clustering. J. Mach. Learn. Res. **4**(Nov), 1001–1037 (2003)
15. Yao, J., Dash, M., Tan, S.T., et al.: Entropy-based fuzzy clustering and fuzzy modeling. Fuzzy Sets Syst. **113**(3), 381–388 (2000)
16. Yang, J.M., Liao, W.C., Wu, W.C., et al.: Trend analysis of machine learning-a text mining and document clustering methodology. In: 2009 International Conference on New Trends in Information and Service Science, pp. 481–486. IEEE (2009)

17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
18. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1981)
19. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2011)
20. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 515–524. ACM (2002)
21. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. Comput. J. **26**(4), 354–359 (1983)
22. U.S. National Science & Technology Council. Emerging technology to support an aging population. USOSTP (2019)