Research on Tourism English Translation System Based on Fuzzy Clustering Algorithm

Jianzhou Cui

Wuxi City College of Vocational Technology, Wuxi, Jiangsu, China

Abstract: The emergence of economic globalization and the Internet have precipitated extensive and profound international exchanges and collaborations. Language barriers have surfaced as the primary impediment to effective international communication and cooperation. This study is dedicated to the development of a tourism English translation system utilizing the fuzzy clustering algorithm. The system's functionalities undergo testing and validation through black box testing. Key performance indicators like response time and throughput are scrutinized to evaluate the system's effectiveness. The performance evaluation involves monitoring specific checkpoints in test cases to ascertain whether the system aligns with the essential performance criteria. The system's response time and throughput take the forefront in this educational system assessment. Use-cases are categorized based on the number of online users, with the client's response time being assessed in each scenario. Successful completion of the criteria deems a use case qualified; conversely, failure designates it as unqualified, with any system deficiencies recorded within the testing framework. Data analysis reveals that the integration of the enhanced PCM algorithm elevates C-FCA's clustering data accuracy to 95%. Consequently, the findings signify that the fuzzy clustering algorithm significantly amplifies the precision of the tourism English translation system.

Keywords: Fuzzy Clustering Algorithm; Tourism English; Cluster Analysis; Translation System

1. Introduction

Recent years witnessed the shift towards prioritizing the development of human personality, creativity, and subjectivity as the central ideology in educational reform and pedagogical experimentation. This initiative has cultivated a shared consensus on the importance of recognizing and highlighting the subjectivity and positioning of students within the educational context [1-2]. By aligning with theoretical principles, the examination of survey findings on undergraduate tourism English activities and student subjectivity highlights two key areas. Firstly, it grants educators and learners a deeper comprehension of the core of tourism English activity instruction [3-4]. Secondly, it enables college English teachers to effectively implement activity-based teaching, enabling students to exercise their subjective initiative and cultivate their own subjectivity [5-6]. As a result, it is crucial to evaluate the likeness and disparity among reference points [7]. In the field of clustering analysis, the proximity of reference points is widely employed as a metric to assess their similarity. The prevailing agreement is that as the distance between reference points decreases, their resemblance increases while the level of divergence diminishes correspondingly, and vice versa [8-9]. Nevertheless, it is important to note that algorithm enhancement only tackles some of the many limitations associated with the algorithm. Additionally, distinct algorithms may produce varying clustering outcomes when applied to the same dataset [10].

The remarkable accomplishments originated from the corpus-driven English translation approach. While statistical techniques bolster the advancement of apt mathematical models for translating natural language with greater durability, the rapid growth of Internet technology provides ample access to bilingual or monolingual text collections, thereby enabling comprehensive training of model parameters. Consequently, the application of statistical machine translation based on extensive corpora has been greatly facilitated. As a sentence's foundation, a template facilitates the extraction of sentence structures with the aid of frame analysis. Moreover, frames embody a broad grammatical structure that disregards subtle grammatical intricacies.

2. Fuzzy Clustering Algorithm and English Translation System

2.1 Fuzzy Clustering Algorithm

Clustering entails the discernment and categorization of entities on the basis of their similarities. In this process, devoid of any explicit prior knowledge guidance, distinguishing and classifying becomes indistinct. Fuzzy relationship serves as a more expansive rendition of conventional relationships, employed to articulate the degree of direct association between a pair of elements. The inherent vectors x_k and x_{kj} generate a group of partitions in the cluster analysis of a specific group of target data objects X.

$$\left.\begin{array}{l} X_1 \cup X_2 \cup \ldots \cup X_c = X \\ X_i \cap X_k = \emptyset, \quad 1 \le i \ne k \le c \\ X_i \ne \emptyset, \quad X_i \ne X, \quad 1 \le i \le c \end{array}\right\} \quad (1)$$

The hard partition space of the data object set X is as follows:

$$M_{hc} = \{U \in R^{cn} \mid \mu_{ik} \in \{0,1\}, \forall i,k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{n} \mu_{ik} < n, \forall i\} \quad (2)$$

Fuzzy segmentation is predicated on the expression of ambiguity across all classifications, thereby facilitating the determination of the degree of uncertainty associated with each individual data point's membership in a specific category. This, in turn, leads to a more refined depiction of the inherent characteristics inherent to the actual data. The FCM algorithm divides a dataset into discernible clusters and leverages them for the assessment of performance. The formulation of the objective function is delineated as follows:

$$J = \sum_{j=1}^{m} \sum_{k=1}^{n} (\mu_{jk})^c (d_{jk})^2 \quad (3)$$

As human cognitive abilities and practical needs progress, real-world situations often display inherent uncertainties, making it extremely challenging to achieve complete accuracy. As a result, traditional inflexible partitioning approaches frequently fail to achieve the best possible results.

2.2 Tourism English Translation System

The preprocessing module undertakes two main functions, namely analyzing the structure of the source language text and conducting syntax analysis. The objective of scrutinizing the textual structure of the source language is to obtain the syntactic framework of the sentence at hand through the application of source language syntactic analysis. The current sentence template is obtained through frame analysis, subsequently undergoing a search and matching process with the existing template library of the target language. Eventually, the target language sentence is generated by modifying the target language template. Since the decoder requires significant storage space, making it resource-intensive, statistical machine translation systems can only perform real-time translation. A typical decoder necessitates approximately 500GB of hard disk space, rendering it impractical for mobile devices. Consequently, only preprocessing, post-processing, and basic translation tasks can be accomplished on mobile devices. The core translation tasks are primarily executed in the cloud.

3. System Test

3.1 Experimental Environment

The experimental test system environment configuration details are depicted in Table 1. Equipped with a Qualcomm Snapdragon S4 Pro 1.5GHz MDM9215 quad-core processor, the Android sampling and recording terminal operates on the Android 4.2 operating system.

Table 1. System Environment

Hardware environment	Processor	Qualcomm Snapdragon S4 Pro 1.5GHZ MDM9215
	RAM	2GB
	External storage	16GB
	Screen size	10.1 inches
Software environment	Terminal	Android 4.2
	System client	Windows 7
	Server	Windows Server 2003
	Database	Microsoft SQL Server 2005

3.2 System Test

The system utilizes the black box testing approach to carry out functional testing and verification. Performance assessment mainly includes factors such as response time and throughput. The conclusion of the system's performance test entails the surveillance of specified checkpoints within the test cases. The significance of performance testing lies in confirming the adherence to the fundamental performance criteria of the system. Emphasis is placed on evaluating the response time and throughput of the educational system. The system's performance test cases are categorized based on the concurrent online user levels, with an examination of the client's response time in each unique scenario.

4. Discussion

4.1 Algorithm Performance Analysis

Each algorithm conducted 50 experiments and recorded the number of accurate classifications for the Iris dataset. Out of the 50 experiments, 3 experiments achieved the optimal cluster numbers, with detailed results presented in Table 2. The analysis of the results reveals that the enhanced algorithm demonstrates superior clustering accuracy in comparison to the traditional Fuzzy C-Means (FCM) algorithm, achieving a precision rate of 94%. Consequently, the enhanced algorithm is deemed efficacious. In terms of the average runtime, the upgraded FCM algorithm outperforms its traditional counterpart. This proficiency can be attributed to the arbitrary selection of initial cluster centers by the traditional FCM algorithm, leading to increased time and cost for convergence. In contrast, the enhanced FCM algorithm utilizes the refined K-means clustering outcomes for the selection of initial cluster centers, often yielding the optimal global solution for datasets with uncomplicated structures. Hence, the enhanced FCM algorithm achieves faster convergence than its predecessor. Moreover, both Random Fourier Features (RFF) and Quasi-Monte Carlo Feature (QMCF) exhibit increasingly precise approximations of the kernel function as the feature dimension expands. Remarkably, QMCF features demonstrate a more accurate approximation effect than RFF features across all feature dimensions. In some instances, even low-dimensional QMCF features outshine high-dimensional RFF features in accurately approximating the kernel function.

Table 2. Experimental Results

Algorithm	Traditional FCM algorithm	Improved algorithm
50 correct classification statistics	44 times	47 times
Correct rate	88%	94%
Operation hours	298s	207s

The outcomes of the IRIS dataset are presented in Figure 1. Concerning the IRIS data, the Fuzzy C-means (FCM) algorithm produces results that generally exhibit a common number of errors and accuracy. On the other hand, the Weighted Fuzzy C-means (WFCM) algorithm tackles the issue of noise sensitivity and consistent clustering by incorporating the notion of potential fuzzy clustering. Although the error score obtained is just 11, it is important to highlight the close proximity of the first two cluster centers to the third cluster center, which deviates from the actual cluster center. Conversely, by leveraging the existing data, the Uncertain Model-Based Partitioning Fuzzy Clustering Algorithm (UMPFCA) identifies uncertainty in the membership relationship among cluster clusters. This approach produces remarkable experimental outcomes, showcasing a high accuracy rate and clustering centers that closely align with the authentic clustering center. Among the four algorithms, the lowest number of errors can be attributed to the Credibility-based Fuzzy Clustering Algorithm (C-FCA) due to the influence of the synergy coefficient. As a result, a higher accuracy rate is achieved, accompanied by clustering centers that bear a striking resemblance to the actual clustering center.

Furthermore, C-FCA builds upon the superior Possibilistic C-means (PCM) algorithm to augment clustering results, leading to a further enhancement in the accuracy of correctly clustered data to 95%. In contrast, the Improved Genetic Algorithm Fuzzy K-Means (IGAFKM) algorithm, despite processing the same volume of data, incorporates an initial center selection process to amplify the global search capability. Unfortunately, this indicates a lengthier processing time. Moreover, the increased weight assigned to distance favors points that are farther away from the selected initial center, potentially encompassing outliers. Consequently, the selection process may disregard truly superior initial centers, thereby adversely impacting algorithm accuracy, recall rate, and overall performance.

Comparative studies have conclusively established that SWFCM triumphs over the alternative segmentation techniques, manifesting its unparalleled excellence in segmentation outcomes. The suggested implementation of the initial value selection technique proposed within this article contributes to a reduction in the iterations demanded. It is imperative to accentuate that the conventional FCM tends to encounter local convergence, which consequently undermines the optimality of the segmentation outcomes. Conversely, due to the unpredictable nature of the initial values, SWFCM necessitates a range of iterations spanning from ten to several hundred. The subsequent table elucidates the average outcomes obtained from an extensive array of experiments.

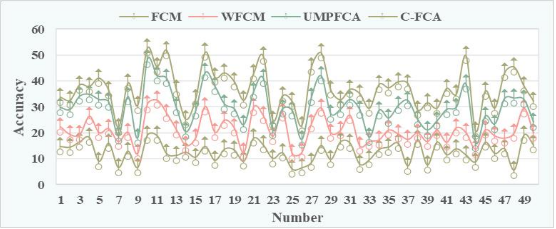

Figure 1. Results of the IRIS Data Set

4.2 System Test Results

The English translation system operates on a personal computer outfitted with a Pentium 4 processor and 256MB of RAM. Throughout its operation, the system utilizes roughly 6MB of memory and has the capacity to process approximately 8000 words per minute. Through evaluations on test sentences, the system displays enhanced proficiency in tackling ambiguity challenges related to sentence and word segmentation, categorical terms, ambiguous structures, and translation choices. Empirical findings indicate that the FCM algorithm is prone to converging to local optima, unlike the other three algorithms that demonstrate more reliable performance. Specifically, within the abscissa range of 10 to 16, the FCM algorithm and the IFCDE1 algorithm yield closely aligned values. However, while the IFCDE1 algorithm continues to converge, the FCM algorithm becomes trapped in a local optimum. As the experimental output represents the minimum value of the objective function across the population, the resulting curve remains monotonically stagnant. The evaluation of the four translation systems involves scoring candidate translations from two perspectives: word-based and word-based. Statistical observations, depicted in Figure 2, unveil substantial shifts in the scores of both BLEU and NIST evaluation methods following word segmentation. Notably, the BLEU score rises while the NIST evaluation score declines. This increase in the BLEU score can be attributed to calculating sentence length based on words as the smallest unit following word segmentation, thereby leading to a reduced denominator in the calculation formula and a higher BLEU score. Under noise-free conditions, the clustering algorithm with membership constraints and dynamic weight adjustment exhibits consistent performance across all three datasets. However, the clustering algorithm featuring dynamic weight adjustment, which loosens membership constraints, generally exhibits inferior performance compared to its statically adjusted weight counterpart. Moreover, the clustering algorithm that incorporates membership constraints and assigns equal weights of 1 to both samples and features outperforms the clustering algorithm that relaxes the membership constraints.

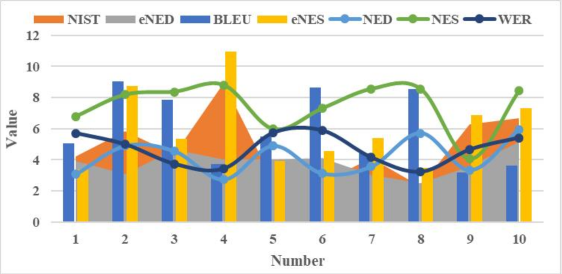

Figure 2. Statistical Results

5. Conclusions

In summary, the experimental comparison presented in this essay demonstrates the superior efficacy of SWFCM in segmentation, surpassing the other two methods. The utilization of the aforementioned technique for initial value selection contributes to a reduction in iterations. However, it is important to acknowledge that standard FCM tends to converge locally, compromising the assurance of optimal segmentation outcomes. Furthermore, due to the random nature of initial values, the number of iterations in SWFCM may exhibit significant variation. Moreover, the translation system for Tourism English demonstrates its excellence by effectively addressing various ambiguity issues, including sentence and word segmentation, categorical words, ambiguity structure, and translation selection. The evaluation scores of BLEU and NIST undergo noticeable changes after word segmentation, with an improvement in BLEU and a decrease in NIST. This variation in BLEU score can be attributed to the computation of sentence length with words as the smallest unit, resulting in a smaller denominator in the calculation formula and ultimately inflating the BLEU score.

Furthermore, the clustering algorithm, incorporating membership constraints and dynamically adjusted weights, demonstrates comparable performance across the three datasets in the absence of noise. However, the clustering algorithm that employs dynamically adjusted weights, relaxing the membership constraints, generally underperforms in comparison to the clustering algorithm with statically adjusted weights. Finally, it should be stressed that the clustering algorithm, when applying membership constraint and setting sample and feature weights to 1, outperforms the clustering algorithm with relaxed membership constraint.

Moving forward, future investigations should prioritize further examination and refinement of SWFCM as a segmentation method, exploration of techniques to alleviate the issue of local convergence in standard FCM, and refinement in the dynamic adjustment of weights in clustering algorithms. Additionally, research endeavors should be undertaken to assess the applicability and scalability of the Tourism English translation system in a broader range of contexts.

References

[1] Feng, R. Xiaoyan, L. Chunhua, et al. "An Improved Collaborative Filtering Recommendation Algorithm Using Singular Value Decomposition and K-means Clustering." Journal of Computer Science and Technology 33.5 (2018): 981-990.

[2] K. Wang, Q. Zhang, Z. Liu. "Object Segmentation in Videos using Adaptive Mean Shift Clustering Based on Fuzzy C-means Algorithm." IEEE Transactions on Image Processing 27.11 (2018): 5432-5445.

[3] M. Chen, Y. Liu, F. Zhang, et al. "A Hybrid Clustering Algorithm for Resource Allocation in Cognitive Radio Networks." Journal of Network and Computer Applications 127 (2019): 162-174.

[4] A. Singh, V. Sharma, N. Kumar, et al. "Enhanced Energy Efficient Clustering Algorithm using Fuzzy Logic in Wireless Sensor Networks." Computers & Electrical Engineering 79 (2019): 106-120.

[5] Sharma, N. Gupta, A. Kumar, et al. "Hybridized Energy Efficient Clustering Protocol using Fuzzy Logic for Wireless Sensor Networks." Journal of King Saud University-Computer and Information Sciences 32.2 (2020): 242-248.

[6] Smith J, Brown K. Application of Hybrid K-means and Particle Swarm Optimization Algorithm for Feature Selection in Data Mining. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING, 2019, 7(3):102-108.

[7] Johnson R, Williams S, Davis P, et al. Optimization of Industrial Production Processes Using Fuzzy Linear Programming and Multi-objective Clustering Algorithm. Mathematical Problems in Engineering, 2019, 2019(6):45-62.

[8] Patel, R. Shah, V. Patel, et al. "Improving Network Security with Fuzzy Attribute Clustering and Machine Learning Techniques." International Journal of Computer Applications 186.17 (2018): 36-41.

[9] K. Patel, S. Gupta, R. Sharma. "Application of Quantum-inspired Optimization for Fuzzy Clustering in Medical Image Segmentation." International Journal of Imaging Systems and Technology 29.4 (2019): 351-359.

[10] A. Das, S. Biswas, R. Saha, et al. "Energy-Efficient Data Fusion Technique for Multi-hop Routing in Wireless Sensor Networks using Hybrid Genetic-Fuzzy Algorithm." International Journal of Ad Hoc and Ubiquitous Computing 32.1 (2020): 58-69.