A Clustering Algorithm Based on Latent Semantic Matrix

Buyu Wang  
_School of Computer Science and Information Technology, Inner Mongolia Agricultural University, \Hohhot010018, Inner Mongolia, China_

Lumeng Chao*  
_School of Computer Science and Information Technology, Inner Mongolia Agricultural University, \Hohhot010018, Inner Mongolia, China_
*Corresponding author (E-mail: clmjet@qq.com)

Abstract  
Establishing a personalized tourism product autonomous push system is a core problem of wisdom tourism. The huge amount of tourism-related data in the Internet provides data support for wisdom tourism. How to obtain and convert from large amounts of data into personalized target data has become a key issue in building wisdom tourism. The paper will study the algorithms and business processes of wisdom tourism push information acquisition and carry out relevant experiments and verification. In order to precisely procure the web tourism information, the paper proposed a clustering algorithm based on latent semantic matrix. The latent semantic matrix of textual length, core segment, core distance and other latent semantic information is established for target search text. Clustering algorithm is important in highlighting the semantic information of different positions, while maintaining the coherence of the semantic information of the retrieval object. It has high accuracy in clustering the retrieval target objects. We first get the latent semantic matrix of tourism information, and use text clustering algorithm on the above matrix, the iterative clustering analysis of tourist information is carried out to find the personalized tourism product model. The model obtains the tourism data from a large amount of Internet data and aggregates and analyzes the data through the relevant algorithm to get the target data. In other industries to solve similar problems, can learn from the algorithm to obtain and process data when the algorithm design and business process arrangements, so as to achieve the purpose of the industry intelligence.

Key words: Personalized Tourism, Wisdom Tourism, Autonomous Push System, Clustering Algorithm, Latent Semantic Matrix.

1. INTRODUCTION  
At present, the core problem of wisdom tourism is how to establish a personalized tourism product autonomous push system (Ulrich Gunter and Irem Önder, 2015). It is one of the hot topics to study the short text tourism information classification method based on large data (Királová and Pavlíčeka, 2015). Mainly through human-computer interaction mode to obtain a limited "travel mode" information (Byun, Kim, Chang and Byun, 2016). It is difficult to provide personalized focus with tens of thousands of "tourist information" of network (Hunter, Chung, Gretzel and Koo, 2015). Therefore, people want to introduce the technology of self-learning without manual guidance and solve this problem by automatically classifying a large number of Chinese tourist information.

The traditional clustering algorithm can not reflect the potential semantics relation between words (Fabrício Olivetti de França, 2016). On the basis of the existing binary classification model, the article (Orhan Kesemen, Özge Tezel and Eda Özkul, 2016) proposed several multi-class classification models based on latent semantics, which solves the problem of multi-genera classification, synonyms and polysemy. However, because of its linear structure based on the classification model, cannot retrieve the target and travel information separately. In article (Frédéric Ros and Serge Guillaume, 2016), a retrieval target was clustered by a matrix representation, which maintains the semantic coherence between the search terms to a certain extent. However, in the construction of sentence-word matrix, the author does not set the semantic center-of-gravity weight, that is, the contribution of target words to classification in different rankings; secondly, unfair matches caused by information source length differences are not considered. That is, the same threshold value, short text and long search target matching easier than the long text and long search target matching. Third, the selection order of the retrieval target is more sensitive, which makes it difficult to guarantee the accuracy rate.

Therefore, by constructing a LSM (latent semantic matrix) based on core word and word distance information, not only maintains the semantic coherence between the word and the word, but also solves the problem of the difference between the retrieval target length and the semantic rank of the semantic order. By
extracting the core segment, the matrix scale is compressed and the same process of alienation is used to minimize the sensitivity of the clustering model to the retrieval order.

2. CORE THESAURUS CONSTRUCTION

Through the analysis of the massive Chinese tourism target information, it can distinguish most of the different scenic areas by the Key_Sets, such as the name, location, route, key scenic spots, scenic features and the mode of target sentences. Therefore, it is possible to improve the accuracy of clustering by assigning a higher weight to key terms in the KeySet.

Definition 1. Extracted from the qualifier, prefix, modification, suffix and statement model of Key_Set to build the title set_KeyWordsLib.

In accordance with the definition, four aspects of construction of the core thesaurus are:
1) Travel mood core thesaurus (KeyMode_Key -WordLib): contains 289 core model terms, such as "tickets", "known as", "covering", "existing" and so on.
2) Place core thesaurus (Place_KeyWordLib): contains 54 locations attribute word, such as "located", "east", "north latitude", "province", "city", "county", "territory", "island" and so on.
3) AttractionsCharismaticCoreThesaurus (Charac_KeyWordLib): contains 104 attractions characteristics related terms, such as "unique", "first", "listed", "first of", "the oldest", "rare", "owned" and so on.
4) Line core thesaurus (Line_KeyWordLib): contains 53 line-related terms, such as "passing", "interchange", "arrival", "ride", "airport" and so on. To the statement "Beijing Forbidden City is located in the central axis of Beijing, is the world's largest existing wooden structure, one of the ancient buildings, has been listed as a world cultural heritage,...") for example, the sentence of the core set of properties have been identified in italics. Through the core vocabulary can measure the importance of the core attributes of the term.

3. LSM MODEL GENERATION

3.1 Auxiliary Model Data Structure Description

To facilitate weight calculation based on the core distance, this chapter introduces the core word-core segment auxiliary model data structure. The structure is a five-element vector (KeyWS Array5), the formal description is as follows: SNum, DNum, KeyWord, KeySentenceNum, Position Respectively, indicate retrieval source number, target number, the core word, sentence number, word position.

CONST KWCount= the core word Number;
TYPE tuple5=ENTRY
SNum,DNum, KeySentenceNum,Position:integer;
KeyWord:empty;
END;
KeyWSArray5= ENTRY data:ARRAY [1.. KWCount] of tuple5p;
END.

3.2 The basic concept of the LSM model

Definition 2. Retrieving target d = {s1, s2,..., sn} , where s_i (i = 1,2,...,n) represents a statement to be retrieved, for the query s_x={w_1,w_2,...,w_m} to be retrieved in target d, if there is a term w_j = {w_1,...,w_m}, s_x = {w_1,w_2,...,w_m}, w_k (k = 1,...,m) in the statement s_x to be retrieved located in the core thesaurus, the term w_j is the core word of the statement s_x, the corresponding order in s_x becomes the position of the core word w_j. If no such w_j exists, then the core word position is the ending order of the statement s_x.d.

Definition 3. The statement s_x={w_1,w_2,...,w_m}, where w_j(1,..., m) denotes a word, if w is the word of the statement s_x, the number of words w_j between any word w_j e s_x and w_i e w and w_j is called core distance.

Corollary 1. Assume that the first word of the statement has a bit order of 0, The order of the core words w is x, the order of any of the other words w_j in the statements except for the core word w is y, the core distance of the term w_j is |y-x|.

Definition 4. Retrieving target collection D = {d_1,d_2,...,d_n}, where d_i (i = 1,..., n) denotes a retrieval target, suppose the length of the search target di is len_d_i, and len_ave_d represents the average of the search target lengths in the set D. If the search target balance factor satisfies a = len_d_i / len_ave_d ≥ 2.5(1 ≤ i ≤ k), through KeyWSArray-5
retrieval target \(d_i\) in addition to the first section and tail section of the core words by the number of paragraphs in descending order, where the former \(\lfloor \alpha \rfloor\) segment is taken as core segment of the retrieving target \(d_i\).

**Definition 5.** Retrieving target collection \(D = \{d_1, d_2, \ldots, d_n\}\), where \(d_i (i=1, 2, \ldots, n)\) represents a search target, a search target \(d_i = \{s_1, \ldots, s_k\}\) consisting of \(k\) statements, the statement \(s_j = \{w_1, \ldots, w_n\}\) consists of \(n\) words and \(w_i \in s_j\), the weight \(w(t_j)\) of the term \(w_i\) in the statement \(s_j\) is:

\[
w(t_j) = tf_j \cdot \log \frac{k}{k_j} \cdot \frac{1}{h} \sum_{j=1}^{n} \frac{\log(len(j))}{dist(i)} \cdot \frac{len_j}{len_{avg}} (1)
\]

where \(tf_j\) denotes the number of occurrences of the word \(w_i\) in the retrieval object \(d_i\), that is, the word frequency of \(w_i\) in \(d_i\), \(k_j\) is the number of search targets in which the term \(w_i\) is present, used to measure the reverse word frequency. \(len(j)\) is the length of the \(s_j\). \(dist(i)\) is core distance of \(w_i\), \(h\) is the number of occurrences of \(w_i\). \(len_{avg}\) represents the length of the search target \(d_i\), \(len_{avg}\) represents the average of the search target lengths in \(D\).

**Theorem 1.** Let \(s_j = \{w_1, w_2, \ldots, w_m\}\) be any of the statements in the retrieval target \(d_i\) and it consists of \(m\) words, \(\hat{w}\) is the core term of the statement \(s_j\), then \(\forall w_j \in s_j \land w_j \in \hat{w}\), the smaller the core distance of the term \(w_i\) in the statement \(s_j\), the greater the weight of the term \(w_i\) is.

**Proof.** For a given statement \(s_j\) and \(w_i\), the formula (1) except \(dist(i)\), the other values are fixed. Only \(dist(i)\) is related to the order distance, so the smaller the \(dist(i)\), the smaller the core distance of the word, the greater the \(w(t_j)\) is.

Theorem 1 shows that for a word in a statement, latent semantic importance is highly correlated with rank relation, and the closer to the core word, the greater its weight, and the greater the contribution to clustering.

**Theorem 2.** If \(D = \{d_1, d_2, \ldots, d_n\}\) represents a set of search targets containing \(n\) search targets, \(\forall d_i, d_j \in D\) and \(d_i \neq d_j\), if \(len_{avg} > len_{d_i}\) then the weight of the search target \(d_i\) is higher than the weight of \(d_j\), the equilibrium factor \(T\) is large.

Theorem 2 is easy to proved by the formula (1), no further proof is given here. Theorem 2 shows that the fairness clustering of different retrieval targets can be achieved by the balance factor when the retrieval target length is large.

### 3.3 LSM model generation

**Definition 6.** Let \(D = \{d_1, d_2, \ldots, d_k\}\) be the set of \(k\) search objects, the set \(D\) is made up of a global set of words \(w = \{w_1, w_2, \ldots, w_m\}\) and the retrieval target \(d_i = \{s_1, \ldots, s_k\} \in D\) consists of \(k\) statements, if the core segment is to be extracted, the dimension reduction processing for the retrieval target \(d_i\) is \(d_i' = \{s_1, \ldots, s_k\} \in d_1 (m' < m)\). Otherwise, the WSM of the target \(d_i\) is the WSM of \(d_i'\) as shown in (3).

\[
M(d_i) = \begin{bmatrix}
w(t_{i1}) & \cdots & w(t_{im}) \\
\vdots & \ddots & \vdots \\
w(t_{ni}) & \cdots & w(t_{nm})
\end{bmatrix} (3)
\]

The LSM set consisting of the set of search objects \(D\) is retrieved:

\[
M(D) = \{M(d_1'), \ldots, M(d_k')\}
\]

Note that the global lexical set \(w = \{w_1, w_2, \ldots, w_m\}\) is the ordered sequence formed after the word segmentation. Each column in the LSM represents a statement, for the value of the word does not appear set to 0. If the \(T\) value of the retrieved object satisfies the requirements of definition 4, the retrieval target can be reduced dimension by extracting the core segment. The LSM constructed in this way is not only a semantically
coherent matrix, but also can effectively highlight the semantic center of gravity of the order, and then retrieve the target balance factor so that there is a difference in the length of the search target can be fairly matched.

4. ANALYSIS AND DESCRIPTION OF ALGORITHM

4.1. Basic Ideology

For any two retrieval targets \( T, R \) in the search target set \( D \), For any two retrieval targets \( d, d' \) in the search target set \( D \), Then \( \pi \) and \( \pi' \) are a partition on search targets \( d \) and \( d' \), respectively. The similarity between the retrieval targets is determined according to the similarity degree thresholds \( \theta \), that is, the similarity between the blocks \( s_i \in \pi \) and \( s_j \in \pi' \). When the overall division ratio of number of similar partitions of two retrieval targets reach the similarity threshold of the retrieval target \( \sigma \), then it is determined that the search targets are similar, and thus the search targets become a cluster.

The statement is equivalent to a column vector of the WSN in the retrieval target. According to matrix multiplication principle, the retrieval target \( d \) is a clustering target search target, and any search target \( d' \) other than \( d \) in the set \( D \) is a candidate retrieval target. The \( sim \) values of the statements and sentences in the two retrieval targets are calculated shown as follow:

\[
sim = \frac{a_i \cdot b_j}{|a_i||b_j|} \quad (4)
\]

\[
f_1 = \frac{\sum w(t_{ij})w(t_{ij})}{\sqrt{\sum w^2(t_{ij}) \sum w^2(t_{ij})}} \quad (5)
\]

\[
f_2 = \frac{\sum w(t_{ij})w(t_{ij})}{\sqrt{\sum w^2(t_{ij}) \sum w^2(t_{ij})}} \quad (6)
\]

\[
f_3 = \frac{\sum w(t_{ij})w(t_{ij})}{\sqrt{\sum w^2(t_{ij}) \sum w^2(t_{ij})}} \quad (7)
\]

\[
f_4 = \frac{\sum w(t_{ij})w(t_{ij})}{\sqrt{\sum w^2(t_{ij}) \sum w^2(t_{ij})}} \quad (8)
\]

\[
(M(d))^T \times (M(d)) = \begin{pmatrix} f_1 & \cdots & f_2 \\ \vdots & \ddots & \vdots \\ f_3 & \cdots & f_4 \end{pmatrix} \quad (9)
\]

In formula (9), if the maximum value of a row is greater than \( \theta \), you can determine the search target corresponding to the two statements are similar. When in formula (9) there are a number of similar sentences, and the ratio of similar sentences to the total number of retrieved target sentences is greater than or equal to \( \sigma \), the two search targets can be considered to belong to the same cluster. The selection of pacesetting targets is a key link to the process of similarity calculation, the order of selection has different effects on the clustering results. The text clustering process can minimize the sensitivity of the clustering results to the retrieval order. The clustering process involved in the algorithm is described as follows:

1) For the search target set \( D = \{d_1, d_2, \ldots, d_k\} \) to be clustered, randomly select a retrieval target \( d_x \in D \) as a model search target, and then cluster with the remaining \( k-1 \) retrieval targets in \( D \) to form a cluster \( \text{cluster}_i = \{d_x, d_j, \cdots, d_x\} \).
2) ∀d'_x ∈ cluster_t and d'_x ≠ d_x as a new model search target, re-clustering to form a new cluster cluster_t, cluster_t, cluster_t, ..., cluster_t, where n<k.
3) Then with d'_x ∈ D as a model produced by the same alienation clustering can be expressed as:
\[\text{cluster} = \text{cluster} \cup \{ \text{cluster}, \text{cluster}, \ldots, \text{cluster} \} \]

4.2. Dynamically Extended Clustering Algorithm Description

The similarity of retrieval target is calculated on the basis of establishing latent semantics, namely LSM model, to achieve the objectives of the different tourist search without manual intervention to guide the classification of the target.

**Inputs:** LSM set M(D), retrieval target similarity threshold σ and statement similarity threshold θ of m search targets.

**Output:** n clusters formed by m search targets clustering.

**Process:**
1) Initialization: com_cluster indicates that a set of clusters has been formed; cand_cluster represents a set of candidate cluster retrieval targets;
2) Randomly selecting a retrieval target M(d_x) ∈ M(D) as a clustering target retrieval matrix, adding it to the com_cluster, adding the remaining m-1 candidate retrieval targets to the cand_cluster,
3) Repeat
4) In cand_cluster randomly select a M(d_x), calculated M = M(d_x)T × M(d_x), The number of \(\text{sim}_{\text{max}} \geq \theta\) in the maximum value of each row in M is counted as the value of count, and then the number of candidate search targets is decreased by 1;
5) if the number of candidate retrieval targets is not 0
   then {
   if \(\text{count} \geq \theta\) rowM(d_x)
   then add M(d_x) to the com_cluster collection, and remove M(d_x) from the cand_cluster, goto(3);
   else goto (3);
   }
6) else ∀M(d_x) ∈ cluster ∧ M(d_x) ≠ M(d_x)

   clustering process is carried out to get the final T as a model corresponding to the cluster as follows:
   \[
   \text{cluster} = \text{cluster} \cup \{ \text{cluster}, \text{cluster}, \ldots, \text{cluster} \};
   \]
7) if cand_cluster ≠ null, and choose a new T as the clustering model to retrieve the target matrix, then add it to com_cluster, the rest of the cand_cluster -1 search targets added to the cand_cluster, the number of candidate retrieval targets = |cand_cluster| - 1.
8) until cand_cluster is null.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1 Data Samples

In order to verify the validity of the proposed semantic LSM, content difference and clustering process on the clustering of tourism information retrieval targets, we choose to experiment under different validation sets: set A of 1500 target retrieval targets is selected from the top 10 classical tourist attractions in Inner Mongolia Autonomous Region. The validation set A is sorted according to the length of the content, and 600 length-difference retrieval targets are selected from the first and the last to form set B.

5.2 Analysis of Results

In the experiment, the relative algorithm in (Michael Reiter, Paolo Rota, Florian Kleber, Markus Diem, Stefanie Groeneveld-Krentz and Michael Dworzak,2016) is implemented to compare with the algorithm in this paper. For the sake of the following description, the algorithm is called contrast algorithm.

**Experiment 1**

With set A as the validation set, the change of clustering accuracy θ, σ is determined by three verification processes Figure 1, Figure2 and Figure 3, and a stable threshold combination is determined as σ = 0.32 in Figure 1.0 = 0.58 in Figure 2.
It can be found from Figure 1, Figure 2 and Figure 3 that the required stability threshold interval is determined by two verification processes Figure 1 and Figure 2. By combining the stability thresholds in Figure 3, we obtain the threshold combination $\theta=0.55$ and $\sigma=0.35$, the verification accuracy of clustering reaches the maximum of 96.8%.

**Experiment 2**

Based on the stable threshold combination determined in experiment 1, we verify the clustering effect of this algorithm on the retrieval target with large content length difference, using set B as the verification set. compare with the contrast algorithm shown in Figure 4.

It can be concluded from Fig. 4 that the accuracy of clustering is significantly reduced for the retrieval targets which content length is different at the same threshold, but the algorithm is relatively stable and the accuracy of clustering is very high, significantly higher than the contrast algorithm. Thus, it is effective to search the target balance factor to achieve a fair match of different length search targets.
Experiment 3

Set A is used as the verification set to verify the effect of the clustering process on the clustering results of the randomly selected retrieval order, and the comparison with the contrast algorithm is shown in Figure 5.

![Figure 5 Correspondence between the number of randomly varying documents and the accuracy of clustering](image)

It can be seen from Fig. 5 that the accuracy of clustering in this algorithm is small, which is higher than that of the contrast algorithm, and it is stable in the process of searching target order by random transformation 8 times. It can be seen that the clustering process can reduce the sensitivity of the clustering results to the retrieval order.

The above three verification experiments show that this algorithm is feasible and effective for Chinese tourism individualized product model cluster analysis. The clustering accuracy of the algorithm is improved obviously compared with the contrast algorithm, which shows that LSM can better represent the semantic similarity between different retrieval targets, and it is feasible to balance the search targets with different content length. The clustering process can effectively reduce the dependence of the clustering results on the retrieval catalog.

6. CONCLUSIONS

The paper proposed an algorithm for clustering based on LSM model for tourism information subject classification. By constructing the core thesaurus, the LSM model based on the core distance information was generated to maximize the latent semantic information of the retrieval target, and the retrieval target balance factor was introduced to make the retrieval target with a big difference in length can match fairly, and the clustering dynamic growth process decreased the sensitivity of the clustering results to the retrieval order, the experiment proves its feasibility and effectiveness.

ACKNOWLEDGEMENTS

This work was supported by No.F020104 and No.F020304.

REFERENCES


Michael Reiter, Paolo Rota, Florian Kleber, Markus Diem, Stefanie Groeneveld-Krentz, Michael Dworzak(2016),”Clustering of cell populations in flow cytometry data using a combination of Gaussian mixtures”,Pattern Recognition,60(16),pp.1029–1040.