Text Similarity Calculation based on Domain Ontology and Concept Clustering

ZhiQiang Zhang[1], Lili Gao[2]

[1]Computer and information engineering college, Tianjin University of Technology and Education, Tianjin, China

[2]Foreign Language college, Tianjin University of Technology and Education, Tianjin, China

Keywords: Semantic Text; Domain Ontology; Concept; Document Clustering; Similarity

Abstract. A swarm intelligence-based web document clustering algorithm is proposed in this paper. The main process of this algorithm is to firstly adopt the vector space modal (VSM) to represent the information of web document. The conventional method is adopted, as eliminating the reduction rule of useless words and feature words to acquire the textual characteristic set, and further the document vector is randomly distributed to a plane. The document is clustered through adopting the swarm intelligence-based web document clustering method. Eventually, the clustering result is collected through adopting recursive algorithm. As the experimental result bespeaks, the swarm intelligence-based web document clustering algorithm has the better clustering characteristics. It is able to completely and accurately cluster web document related to the subject. Additionally through comparative analysis from multiple aspects, the clustering results of swarm intelligence-based web document clustering algorithm shall be superior to the clustering result acquired by SOM algorithm.

Introduction

Web has been rapidly developed as the global information space containing great capacity of information and being widely distributed on the globe. The web information retrieval is accordingly becoming a research field attached increasing importance to. Web document clustering is perceived as an important issue in the field of web information retrieval. The document clustering is recognized as the document classification without guidance. Its objective is to subdivide a document set into multiple clusters. Members of the cluster are majorly similar, and the documents among different clusters are less similar to each other. The web document clustering is to cluster the web document in accordance with the contents. It shall merely be adopted to effectively organize the web document but to form the classification module for web document classification.

The document clustering algorithms being currently adopted contains the agglomerative hierarchical clustering, AHC, represented by G-HAC, the document clustering method adopted in Soina system of digital library system of Stanford library, the planar subdivision represented by K-Means algorithm, and the Self-Organizing Maps (SOM) method [1,2,3,4,12] represented by WEBSOM. The first two algorithms manifest the application of classical clustering algorithm in the document clustering. In the meantime, these algorithms carry the defects of being the clustering algorithm. Some of these defects are even more prominent in the document clustering. For instance, AHC algorithm may generate several clusters having no connection with each other and collecting them as the cluster set, which is extremely not conducive to the document clustering. As for K-Means algorithm, besides to pre-establish the number of cluster set K, it is sensitive towards the noise, outliers and input sequence. Such algorithm shall face the difficulty to pre-establish the appropriate cluster set number while clustering a document. Additionally, its sensitivity towards the outliers shall affect the quality of document clustering. The SOM is advantageous in surmounting the defects as mentioned, and yet this algorithm commonly requires multiple-layer clusters while clustering the document, viz. on the basis of last-time clustering result, the larger cluster shall be carried out the SOM clustering once again as to acquire the multi-layer clustering results, as the

multi-layer SOM document clustering algorithm proposed by Hsinchun Chen.

Swarm Intelligence and Relevant Clustering Algorithms

The swarm intelligence emerged massively in the wake of group behaviors of social animals has aroused increasing attention from the public. The algorithms inspired by the foraging, sweeping lair and other behaviors carried out by social animals to resolve the practical problems receive the remarkable success. The application examples of these algorithms in the fields of combinational optimization, communication network and robot are increased exponentially [5,6,7,8,9]. Bonabeau, et al. perceive the swarm intelligence as the algorithm and distributed problem solving equipment inspired by the collective behaviors of social insect group and other animal groups [10]. The swarm intelligence has the characteristic that the minimum intelligent but autonomous individual is able to realize the completely distributed control through drawing on the interactions among individuals and between individual and environment. Additionally, the swarm intelligence is characterized by the self-organization, extendibility and robustness.

The swarm intelligence-based web document clustering algorithm is originated from the classified study of ant colony and ant egg. E. Lumer and B. Faieta adopt the formicary classification model proposed by Deneubourg in the analysis of data clustering[7]. Wu Bin et al. propose the Clustering algorithm based on Swarm Intelligence (CSI) on the basis of simplifying the classification model, and successfully adopt this algorithm in the clustering analysis of standard machine learning database and the customer behavioral analysis of customer relation management (CRM) [14,15]. Compared with the classical hierarchical clustering algorithm and K mean value dynamic clustering algorithm, the CSI has the common characteristics of the swarm intelligent algorithm. It draws on the interactions among individuals and between individual and environment to realize the self-organized clustering without requiring the pre-establishing of clustering center number. It is characterized by the robustness and visualization. The main thought of CSI is to randomly place the objects to be clustered in the 2D grid environment. Each object has a random initial position. Every ant is able to move on the grid, and the group similarity of the current objects in the partial environment is measured. The function is transformed in accordance with the probability, and the group similarity is transformed to the probability of mobile objects. The objects shall be elevated or dropped abiding by such probability. The collective action of ant group makes the objects of the same category collected in the same spatial region. In the following paper, the definitions of two significant concepts, the group similarity and the probability-transformed function shall be provided.

Definition 1: The group similarity is deemed as the comprehensive similarity between the model (object) to be clustered and all other models in the certain partial environment.

Formula (1) is the basic measuring formula of group similarity.

$$f(o_i) = \sum_{Oj \in Neigh(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha} \right] \tag{1}$$

Where $Neigh(r)$ refers to the partial environment. It is commonly referred as the round area with the radial of r in the 2D grid environment.

Definition 2: The probability-transformed function is to transfer the group similarity as the probability of simple individual movement model (object) to be cultured.

It is the function with the variable of group similarity. The range of this function is [0,1]. In the meantime, the probability-transformed function can also be recognized as the probability-transformed curve. It is commonly combined by two relative curves respectively corresponded to the picking-up transformation probability of model p_p and the dropping transformation probability of model p_d . The main principle to formulate the probability-transformed function is that, the picking-up transformation probability of model shall be smaller with the increase of group similarity, and the picking-up transformation probability of model shall be larger with the decrease of group similarity. The dropping transformation probability

of model basically abides by the converse regulation.

In accordance with the main principle to formulate the probability-transformed function, CSI shall simplify the quadratic cure proposed by Deneubourg as the line with the slope of k, as shown in formula (2) and (3).

$$Pp = \begin{cases} 1 & f(o_i) \le 0 \\ 1 - k \times f(o_i) & 0 < f(o_i) \le 1/k \\ 0 & f(o_i) > 1/k \end{cases} \tag{2}$$

$$Pd = \begin{cases} 1 & f(o_i) \ge 1/k \\ k \times f(o_i) & 0 < f(o_i) < 1/k \\ 0 & f(o_i) \le 0 \end{cases} \tag{3}$$

Swarm Intelligence-based Web Document Clustering Algorithm

The Swarm Intelligence-based Web Document Clustering Algorithm is the application of CSI algorithm in the Web document clustering. Its main process is to initially represent the Web document information through adopting VSM, eliminate the reduction rule of stop words and feature words through adopting the conventional method to acquire the feature textual set, further randomly distribute the document vector to the plane, cluster the document through adopting the clustering method based on the swarm intelligence, and eventually collect the clustering result on the plane through adopting recursive algorithm.

Given that the VSM is the most extensively adopted model in information retrieval, the algorithm hereof also adopts this model. The feature word and its weight are adopted by the VSM to represent the document information. The vector $d = (w_1, w_2, w_3, \cdots, w_m)$ refers to the feature word and its weight of document *d*, where *m* refers to all numbers of words in the document set, $w_i (i = 1, \cdots, m)$ refers to the weight of word t_i in the document *d*. to acquire the feature words, the words of Web Chinese document shall be initially segmented. The table of stop words is adopted to get rid of the stop words from the document feature set. Furthermore, the IDF is adopted to reduce the feature words. The feature words within the certain limits of IDF in the document set shall be preserved as the document feature set. The TF is selected as the weight of feature words. Hence, a set of vectors representing the document set is acquired, which is also the model set to be clustered.

The similarity function of web document vector adopts the cosine function being commonly used in IR field. The return value of function lies between [0,1]. The similarity shall be larger with the increase of value. The calculation function is shown in formula (4).

$$sim(d_i, d_j) = \frac{w_{i1} * w_{j1} + \cdots + w_{im} * w_{jm}}{|d_i| * |d_j|} \tag{4}$$

where $(w_{i1}, \cdots, w_{im})$ refers to the feature vector of document d_i, and $(w_{j1}, \cdots, w_{jm})$ refers to the feature vector of document d_j.

$$|d_i| = \sqrt{\sum_{k=1}^{m} w_{ik}^2}, \quad |d_j| = \sqrt{\sum_{k=1}^{m} w_{jk}^2}.$$

The group similarity coefficient α in the formula (1) is the key coefficient to measure the group similarity. To easily select α, the basic measuring function of group similarity is rectified as follow:

$$f(d_i) = \sum_{Oj \in Neigh(r)} \left[1 - \frac{10 * (1 - sim(d_i, d_j))}{\alpha} \right] \tag{5}$$

Hence, the α can be valued between [1,10].

The swarm intelligence-based web document clustering algorithm is described as follow:

Step 1, initialize the parameter, including α, $ant\_number, k, R, size,$ maximum cycle index n, annotated category value *clusterno*, etc.

Step 2, randomly distribute the model to be clustered on the plane, viz. randomly endow every model with a (x,y) coordinate.

Step 3, endow the initial model value for one group of ant, the non-loaded condition is initialized.

Step 4, *for* $i=1,2\ldots\ldots n$,

Step 4.1, *for* $j=1,2\ldots ant\_number$

Step 4.1.1, select the coordinate of this ant in the initialize condition as the center, r as the observation radius, the group similarity within the observation radius of this model is calculated through adopting formula (5).

Step 4.1.2, If this ant is non-loaded, the formula (2) is to be adopted to calculate the picking-up probability Pp.

Step 4.1.3, compared with the random probability of Pr. If Pp<Pr, the ant shall not pick up such model, and the ant is randomly endowed with a model value. Otherwise, the ant shall pick up the module, the condition of ant shall be modified as being loaded, and the ant is randomly endowed with a new coordinate.

Step 4.1.4, if this any is loaded, the formula (3) is adopted to calculate the dropping probability Pd.

Step 4.1.5, compared with the random probability Pr. If Pd>Pr, the ant shall drop the model. The model is endowed with the coordinate of the ant. The condition of ant is modified as non-loaded. A model value is randomly endowed for the ant. Otherwise the ant shall continue to carry this model, with the condition of loaded. A new coordinate shall be randomly endowed for ant once again.

Step 5, *for* $i=1,2\ldots\ldots pattern\_num$; *// for every model*

Step 5.1, if this model category is not annotated

Step 5.1.1, annotate the category of this model

Step 5.1.2, adopt the annotated value of the same category to recurve all the models with the distance less than $dist$, viz. collect all the models belonging to the same cluster on the plane

Step 5.1.3, annotated value of category clusterno++;

It can be acquired that steps 1-3 are the initial phase of algorithm. Its mainly function is to carry out the program initialization and the random distribution model on plane; step 4 is the clustering course based on the swarm intelligence; step 5 is the course to annotate the category, which is the course to collect the clustering result. The main step of this algorithm is step 4. Under the premise of adopting the similar matrix, the rough analysis of its complexity is $O(n \cdot ant\_number \cdot (Aver + R^2))$, where n refers to the pre-established cycle index, ant_number refers to the number of ant, $Aver$ refers to the average model number of the partial environment, and Я refers to the observation radius. As Я is commonly small, the rough analysis of algorithm's time complexity is $O(n \cdot ant\_number \cdot Aver)$.

Conclusion

As the swarm intelligence-based web document clustering algorithm is adopted, the content of each cluster can be further annotated, the low-level query catalog of the entire document set can be formed, as to more effectively organize the web document; additionally the method to acquire the average weight can be adopted to acquire the template vector of clustering model as to be used in the classification of scatter diagram and new document. The clustering algorithm application in the rational selection of parameter, clustering efficiency elevation, and large-scale web document clustering can be further delved into.

Acknowledgement

Research and development fund of Tianjin vocational and Technical Normal University (kj10-16).

References

[1] Yingyue Zhang, Jennifer W. Chan, Alysha Moretti, and Kathryn E. Uhrich, Designing Polymers with Sugar-based Advantages for Bioactive Delivery Applications, Journal of Controlled Release, 2015, 219, 355-368.

[2] Yingyue Zhang, Qi Li, William J. Welsh, Prabhas V. Moghe, and Kathryn E. Uhrich, Micellar and Structural Stability of Nanoscale Amphiphilic Polymers: Implications for Anti-atherosclerotic Bioactivity, Biomaterials, 2016, 84, 230-240.

[3] Jennifer W. Chan, Yingyue Zhang, and Kathryn E. Uhrich, Amphiphilic Macromolecule Self-Assembled Monolayers Suppress Smooth Muscle Cell Proliferation, Bioconjugate Chemistry, 2015, 26(7), 1359-1369.

[4] Yingyue Zhang, Evan Mintzer, and Kathryn E. Uhrich, Synthesis and Characterization of PEGylated Bolaamphiphiles with Enhanced Retention in Liposomes, Journal of Colloid and Interface Science, 2016, 482, 19-26.

[5] Yingyue Zhang, Ammar Algburi, Ning Wang, Vladyslav Kholodovych, Drym O. Oh, Michael Chikindas, and Kathryn E. Uhrich, Self-assembled Cationic Amphiphiles as Antimicrobial Peptides Mimics: Role of Hydrophobicity, Linkage Type, and Assembly State, Nanomedicine: Nanotechnology, Biology and Medicine, 2017, 13(2), 343-352.

[6] Jonathan J. Faig, Alysha Moretti, Laurie B. Joseph, Yingyue Zhang, Mary Joy Nova, Kervin Smith, and Kathryn E. Uhrich, Biodegradable Kojic Acid-Based Polymers: Controlled Delivery of Bioactives for Melanogenesis Inhibition, Biomacromolecules, 2017, 18(2), 363-373.