Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 1
AIEA 2023

Application and Research of Multi-Feature Fusion Tag Propagation Computer Algorithm in Image Search and Matching

Jiale Li

Ningbo ladder Education Technology Co., Ltd, Ningbo 315000, China

li.jiale@163.com

Abstract. When analyzing the advantages and disadvantages of common community discovery algorithms, the paper points out that the label propagation algorithm (LPA) has low time complexity, does not need to set the number of communities in advance, and the calculation process is simple. When dealing with large and complex networks, it has high the characteristics of efficiency. However, the algorithm does not consider the similarity of adjacent nodes in the network structure and content in the process of label propagation. Therefore, from the perspective of node similarity, the paper proposes a multi-feature fusion label propagation algorithm. The algorithm first uses the Sim Rank algorithm to calculate the structural similarity of the nodes in the network, and at the same time uses the main body model to obtain the topic distribution of the node content, and calculates the similarity of the topic distribution of different nodes, and finally merges the two similarities to be the label propagated by adjacent nodes, Give the corresponding weight to improve the communication strategy. Experimental comparison shows that this algorithm is better than the traditional label propagation algorithm.

Keywords: Community Discovery; LPA; SimRank; Topic Model.

1. Introduction

As the scale of social networking sites such as Weibo continues to grow, individual users hope to quickly, accurately and efficiently find users and groups with common hobbies that are consistent with their own body interests, so that users can communicate with each other, thereby giving birth to more Enlightenment and resonance. However, in the face of massive amounts of information, it is impossible to accurately and timely find like-minded friends only through search. Therefore, it is necessary to mine user information in the Weibo network to find out user communities with similar interests. Therefore, for social networks, the community discovery technology can dig out users' potential interest communities, study the relationship structure of user communities, and have important significance for interest recommendation and accurate advertising. Commonly used community discovery algorithms, such as the CPM algorithm has high time complexity and complex calculations, and are not suitable for community discovery on social networks; another example is the algorithm based on hierarchical clustering that needs to determine the number of communities in advance, and it is not suitable for social networks. Community discovery. The label propagation algorithm (LPA) algorithm has low time complexity, no need to set the number of communities in advance, simple calculation process, and high efficiency when dealing with large and complex networks, so it is widely used in the analysis and research of community discovery in [1].

The SimRank algorithm measures the similarity of nodes in the network structure, and the topic model can measure the topic distribution of the node content. Inspired by this, this chapter starts from the perspective of node similarity, and uses the SimRank algorithm to measure the similarity between nodes in the network structure based on the label propagation algorithm. At the same time, it combines the similarity of the topic distribution of the node content to improve the propagation strategy to reduce the experimental results. Randomness.

2. Background Knowledge Introduction

2.1. LDA Topic Model

The topic model is a more classic generative probability model in the field of text processing. The main idea of the model is: a document can be represented by several topics in a certain probability distribution form, and each topic can be represented by several words in a certain probability distribution form, and the documents and words are related to each other through topics. The LDA generation process is shown in Figure1. Maintaining the Integrity of the Specifications.

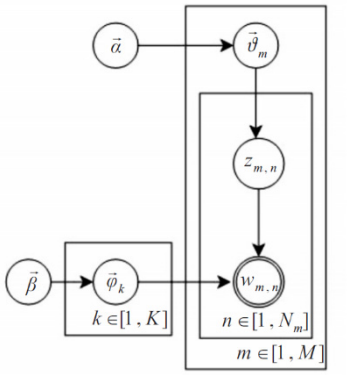

Figure 1. LDA probability graph model.

In Figure 1 the double circle part represents the observable variable, the single circle part represents the unobservable variable, the arrow represents the conditional probability dependence between the two variables, the box represents repeated sampling, and the capital letters in the box represent the number of samples. Taking the mth document as an example, the generation process of LDA is as follows:

1) First generate Nm words in the document according to the Poisson distribution; 2) Use the Dirichlet distribution with the parameter to generate the topic distribution of the document $\vec{\vartheta}_m \sim Dir(\vec{\alpha})$; 3) Use the Dirichlet distribution with the parameter β to generate the topic word distribution $\vec{\varphi}_k \sim Dir(\vec{\beta}), k = 1,2,...,k;$ 4) Pair Each word $w_{m,n}, n = 1.2,..., N_m$ under the document: (1) Sampling from the document topic polynomial probability distribution of the mth document to generate a topic $z_{m,n} \sim Multi(\vec{\vartheta}_m), z_{m,n} \in \{1,2,...,K\}$. (2) Sampling from the polynomial probability distribution of topic words with topic $z_{m,n}$ to generate a word, $w_{m,n} \sim Multi(\overrightarrow{\varphi_{z_{m,n}}}), w_{m,n} \in \{w_1, w_2, ..., w_v\}$. 5) The above operation is the process of generating the topic model. The solution of the LDA model is the process of reasoning about the parameters of the LDA. Gibbs sampling [2] is commonly used to solve the problem. This method finally obtains the estimated value of the target parameter according to the given optimization objective function.

2.2. SimRank Algorithm

Define directed graph $G = (V, E)$, where V is the set of nodes in the graph, E is the set of edges in the graph, and the edge represents the relationship between two nodes. For any node a in the graph, use $I(a)$ to represent the edge point to it. Corresponding vertex set, $|I(a)|$ represents the number of vertices contained in the vertex set. The main idea of the algorithm is that the similarity of two nodes in the network depends on the similarity of the set of two adjacent nodes in the network. For any two vertices a and b $s(a,b)$ in the graph, which represents the similarity between vertex a and vertex b, SimRank calculates the similarity $s(a,b)$ between the two vertices, as shown in formula (1), where F is the attenuation factor, usually a constant.

$$s(a.b)=\begin{cases}1a=b\\ \dfrac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}s(I_1(a)I_1(b))I(a)and(b)\neq\varnothing\\ 0othersize\end{cases}\tag{1}$$

The Sim Rank algorithm usually uses an iterative calculation method to calculate the similarity. The similarity between vertex a and vertex b after the kth iteration is represented by $s^k(a,b)$, and the similarity value of the k+1 vertex can be calculated by the k-th vertex similarity value. (2) As follows:

$$s^{k+1}(a.b)=\begin{cases}1a=b\\ \dfrac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}s(I_1(a)I_1(b))I(a)and(b)\neq\varnothing\\ 0othersize\end{cases}\tag{2}$$

The initial state of the iteration is shown in formula (3):

$$s^0(a,b)=\begin{cases}1a=b\\ 0a\neq b\end{cases}\tag{3}$$

When k approaches infinity, the similarity $\lim_{k\to a}s^k(a,b)=s(a,b)$ between vertices is obtained. It is usually necessary to set a threshold ε, which is used to indicate the accuracy, and the iteration will stop when the accuracy is reached after the algorithm iteration.

2.3. Label Propagation Algorithm

The core idea of label propagation is: first assign a unique label to each node in the network, which represents the label of the community to which the current node belongs, and then count the number of labels of the nodes connected to the current node, and set the largest number of labels The label is assigned to the current node. After multiple iterations, the label of each node in the network becomes stable. At this time, nodes with the same label form a community. The update operation rules are as formula (5):

$$l_i=\arg\max_{l}\sum_{k\in N(v_i)}\delta(l_{k,}l)\tag{4}$$

Where li is the label of the target node i waiting to be updated; $N(v_i)$ is the set of adjacent nodes of node vi; k is one of the adjacent nodes; $l_{k,}$ is the label of node k; $\delta(l_{k,}l)$ represents the Kronecker function, and its input variables are two Integer, when two integers are equal, the output value of the function is 1, otherwise the output value is 0. The optimal solution of this formula is the label with the largest number of labels among the adjacent nodes of node v.

The specific algorithm steps are as follows:

1) Initialization: All nodes in the network are assigned a label. In the specific algorithm, the ID of the node is generally assigned to this node as a label, which represents the number of the community to which the node belongs; 2) Update label: For all nodes in the network The node updates the label. The label update rule is to count the frequency of labels appearing in adjacent nodes, and update the label of the current node to the label with the largest frequency. If there are many tags with the largest frequency, randomly select a tag from the tag with the largest frequency to update as the label of the node; 3) Repeat the above operations until the algorithm stop condition is reached; 4) Community division: Statistics The label of each node, nodes with the same label are in the same community. In this algorithm, the time complexity of assigning labels to each node is O(n), and each iteration takes O(m) to propagate, so the total time complexity is O(m+n). However, in general, the algorithm requires multiple iterations to stabilize the label of the node, but even with multiple iterations, the time complexity is still linear. The time complexity of the algorithm is very small, and it is suitable

for large networks [3].

3. The Algorithm of this Paper

If $G(V,E)$ is a community network, where V is a collection of community nodes, and E is various associations between nodes. For any community node $v_i \in V$, it means its label means the topic θ_i distribution on the content of community node vi, n represents the number of network nodes, and m represents the number of network edges.

In the traditional label propagation algorithm, when a node updates its label, the importance of each neighbor of the node is equally important. As a result, when a node updates its label, the number of labels of the same type becomes the only metric. Inspired by this, when predicting node labels, this paper considers both the similarity between nodes in the network topology graph and the similarity of node content. The higher the similarity, the stronger its influence and the greater the weight; During the update process, the label of a node is no longer uniquely determined by the number of labels in adjacent nodes, but is determined by the weighted sum of labels of adjacent nodes. First consider the similarity between adjacent nodes and the current node in the network topology. Using the SimRank algorithm, the similarity formula $ssim(v_i, v_j)$ of the node structure in the network structure graph G can be obtained, and the similarity of the two nodes in any network structure can be obtained by calculation, and the similarity matrix $SSIM_{n*n}$ can be obtained [4].

$$ssim(v_i, v_j) = \begin{cases} 1 v_i = v_j \\ \dfrac{C}{|I(v_i)||I(v_j)|} \sum_{m=1}^{|I(v_i)|} s(I_m(v_i) I_n(v_j)) I(v_i) and I(v_j) = \varnothing \\ 0 otherwise \end{cases} \tag{5}$$

At the same time, traditional label propagation algorithms often ignore the similarity of node content in the network. The higher the node content similarity, the stronger its label propagation ability. In this paper, the text content of the node is modeled through the topic model, and the topic distribution θ_i of the content of the node vi in the network is obtained, so that the content similarity between nodes can be obtained by calculating the similarity between the topic distributions of the nodes. KL divergence is an important indicator for calculating the dissimilarity between any two probability distributions p and q, as shown in equation (7).

$$KL(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \tag{6}$$

In this paper, KL divergence is used to measure the similarity of node content, that is, the closer the topic distribution of the content on the node, the higher the similarity value. Therefore, the above-mentioned KL divergence is modified and used to calculate the content similarity of nodes, as shown in equation (8).

$$content(v_i, v_j) = 1 - KL(\theta_i, \theta_j) \tag{7}$$

Where θ_i represents the topic distribution of content on node v_i, and θ_j represents the topic distribution of content on node v_j. Finally, the similarity of node v_i, v_j can be obtained by integrating node content similarity, node structure similarity and node structure similarity, as shown in equation (9).

$$sim(v_i, v_j) = ssim(v_i, v_j)^* conten(v_i, v_j) \tag{8}$$

From the perspective of node structure and content similarity, combining the Sim Rank algorithm and topic model, this paper proposes a multi-feature fusion label propagation algorithm. When the algorithm updates the label, it assigns different weights to the labels of adjacent nodes. The final

node's label is determined by Corresponding to the weighted sum of tag weights to determine, the specific formula (10) is as follows.

$$l_i = \arg\max_i \sum_{v_j \in N(v_i)} \delta(l_j, l) \sin(v_j, v_i) \tag{9}$$

Where $\sin(v_j, v_i)$ is the similarity between node v_j and node v_i.

4. Experimental Results and Analysis

In order to evaluate the effectiveness of the multi-feature fusion tag propagation algorithm proposed in this chapter, combined with the distributed crawler program written by Sina API, we crawled 800 users from Sina Weibo, 20,042 Weibo records, and passed The user's friend relationship builds a user's social network and conducts a comparative experiment with the standard tag propagation algorithm [5].

During the experiment, the initial number of communities is the topic number setting of the topic model, and the initial label of the network node is the topic corresponding to the maximum user topic distribution on the network node, that is, the user's interest label. The experiment uses modularity as an indicator to measure the quality of the algorithm. Modularity measures the difference between the distribution of network nodes under complete randomization and the distribution of the detected community structure. The larger the difference, the less randomized the results obtained and the more effective the community detection results. Its formula (11) is as follows:

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \tag{10}$$

Among them, m is the number of entire communities detected by the algorithm; L is the total number of connections of edges in the network structure; ls is the number of connections of edges within the community; d_s is the sum of the degrees of all nodes in the community [6].

In the image retrieval process, it is necessary to effectively evaluate the performance of the algorithm. Recall and Precision are currently the most widely used evaluation criteria in image retrieval. Recall rate refers to the proportion of the number of correlations returned by the system to the number of correlations in the image database, and the precision rate refers to the proportion of the number of correlations returned by the system to the number of all returned images in a retrieval process. Generally speaking, recall rate The higher the sum precision rate, the better the retrieval system, but the two values present a contradictory relationship in general, and there are generally certain conditions to maximize the two values. In order to make the algorithm performance comparison more accurate, the experiment in this paper adopts precision and comprehensive evaluation index F_Measure, where F_Measure is the balance point between recall and precision

$$p(F\_Measure) = \frac{2 \cdot p(\mathrm{Re}\,call) \cdot p(\mathrm{Pr}\,ecision)}{p(\mathrm{Re}\,call) + p(\mathrm{Pr}\,ecision)} \tag{11}$$

It can be seen from the above formula that F_Measure is a comprehensive performance index based on recall and precision. In addition, this article adds time to describe the complexity of the algorithm.

Table 1. Feature search results

Search performance	feature		
	Color characteristics	Texture feature	Comprehensive characteristics
Precision (%)	52.5	50	62.5
F-Measure (%)	46.67	44.44	55.56
Time(/s)	0.936	0.642	1.638

In order to test the effectiveness of this algorithm in comprehensive feature extraction, this paper uses

color and texture features to perform single feature retrieval and comprehensive feature retrieval experiments. The classification algorithm uses the nearest neighbor classification (NN) retrieval algorithm, and the retrieved image is the Corel image database "Food" in the, counts the results of the 40 retrieved images returned, see Table 1.

It can be seen from the experiment in Table 1 that in the retrieval experiment using single feature and comprehensive feature, the retrieval effect of using comprehensive feature is more obvious and better. This shows that the comprehensive feature of color and texture used in this article is used for retrieval. It has retrieval advantages, but the retrieval time will increase due to the increase of the feature dimension when using comprehensive features for retrieval [7].

At the same time, in order to test the feasibility of the R_GLCM_SRC classification algorithm, this paper compares with the classic nearest neighbor classification (NN) retrieval algorithm and support vector machine SVM. The image features adopt the comprehensive features of the color moment and gray-level co-occurrence matrix of this paper. The images are various types of image data from the Corel image database, and the retrieval results of 40 images returned by various types of images are counted. See Tables 2 to 3.

Table 2. NN classification algorithm results

category	Algorithm NN		
	Precision (%)	F-Measure (%)	Time(/s)
character	50	44.44	1.607
beach	55	48.89	1.685
building	35	31.11	1.654
car	60	53.33	1.638
dinosaur	100	88.89	1.636
Elephant	60	53.33	1.653
flower	85	75.56	1.638
horse	77.5	68.89	1.606
mountain	37.5	33.33	1.638
food	62.5	55.56	1.638

Table 3. R_GLCM_SRC classification algorithm results

category	Algorithm R_GLCM_SRC		
	Precision (%)	F-Measure (%)	Time(/s)
character	100	88.89	6.489
beach	100	88.89	5.975
building	100	88.89	5.085
car	100	88.89	5.492
dinosaur	100	88.89	5.429
Elephant	100	88.89	4.868
flower	100	88.89	5.601
horse	100	88.89	4.868
mountain	100	88.89	5.148
food	100	88.89	5.242

The R_GLCM_SRC algorithm has better performance than the nearest neighbor retrieval algorithm and the SVM algorithm. In the process of increasing return images, R_GLCM_SRC is more excellent and more stable in recall and F-Measure comprehensive indicators, which also shows that the R_GLCM_SRC algorithm is feasible in image retrieval [8].

In addition, R_GLCM_SRC takes longer, which is mainly due to the time consumption of the sparse solution process, but this does not affect the retrieval performance. SVM classification algorithm results are shown in Table 4.

Table 4. SVM classification algorithm results

category	Algorithm SVM		
	Precision (%)	F-Measure (%)	Time(/s)
character	57.5	51.11	1.08
beach	70	62.22	1.08
building	47.5	42.22	1.09
car	67.5	60	1.09
dinosaur	100	88.9	1.09
Elephant	60	53.33	1.11
flower	87.5	77.78	1.12
horse	87.5	77.78	1.11
mountain	50	44.44	1.13
food	82.5	73.33	1.13

In order to evaluate the effectiveness of the R_GLCM_SRC algorithm for image retrieval, this paper sets up multiple sets of comparative experiments, which are respectively compared with the image retrieval algorithm based on local binary mode (denoted as R_LBP) and the binary tree complex wavelet image retrieval algorithm using spectral features as R_DT_CWT) conducted an experimental comparison and comparison. This article carried out the retrieval feedback of the "horse", and a total of 25 images were returned. The experimental results are shown in Figure 2.

Figure 2. Search results of various algorithms.

Taking into account the issue of universality, this article sequentially retrieved 10 types of images of people, beaches, buildings, cars, dinosaurs, elephants, flowers, horses, mountains and food, and calculated the accuracy of the returned 25 images, according to " It can be seen that the image retrieval algorithm R_GLCM_SRC has a higher precision and recall rate, and the retrieval performance is more stable. This aspect is due to the integration of color in image feature extraction. And texture features, eliminating more conducive to eliminate the "semantic gap" phenomenon, on the other hand, in the sparse classification algorithm, the calculation of the average residual makes the algorithm more convergent, as shown in Figure 3.

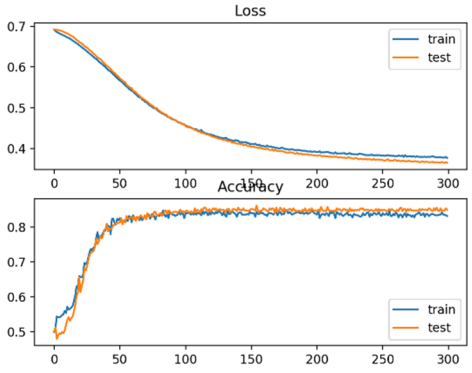

Figure 3. Recall results of each algorithm.

Experiments were carried out to verify the accuracy of the detection community of the algorithm and the stability of the algorithm. The accuracy of the algorithm was compared with the mean value of modularity; the variance of the modularity of the algorithm stability was compared. This experiment was run 10 times on the two algorithms. The experimental results are shown in Figure 4.

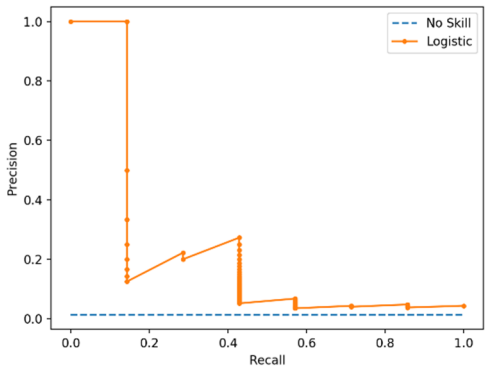

Figure 4. Algorithm comparison chart.

As can be seen in Figure 4, the fusion multi-feature label propagation algorithm proposed in this chapter is significantly higher than the traditional label propagation algorithm in terms of modularity; it is lower than the traditional label propagation algorithm in terms of modularity variance. Experiments show that the multi-feature fusion label propagation algorithm proposed in this paper has improved the accuracy and stability of the algorithm.

5. Conclusion

This paper fully considers the influence of the similarity of the structure and content between nodes on the label propagation in the process of label propagation. First, the SimRank algorithm is used to calculate the structural similarity of the label, and then the topic model is used to obtain the topic distribution of the content on the node. In this way, the similarity of the node content is calculated, and finally the similarity of the two types of information is merged to improve the communication strategy of the label. This algorithm weakens the influence of nodes with different structures and dissimilar contents on label propagation, and at the same time strengthens the influence of nodes with the same structure and similar contents on label propagation, thereby effectively promoting the formation of communities. Experiments show that the algorithm proposed in this paper has better accuracy and stability than the original label propagation algorithm.

Acknowledgments

This work was supported by the science and technology projects in Ningbo high-tech Zone, China (Grant: 2020CX050002).

References

[1] Shi, H. Zhu, H. Wang, J. Yu, S. Y. & Fu, Z. F. (2016). Segment-based adaptive window and multi-feature fusion for stereo matching. Journal of Algorithms & Computational Technology, 10(1), 184–200.

[2] Ren, D. (2020). Research and analysis on precise matching method for multi-feature of fuzzy digital image. International Journal of Computers & Applications, 42(2), 141-149.

[3] Li, & Xirong. (2017). Multimedia systems (accepted) (will be inserted by the editor) tag relevance fusion for social image retrieval. Multimedia Systems, 23(1), 29-40.

[4] Wang, L. J. Han, J. Zhang, Y, & Bai, L. F. (2016). Image fusion via feature residual and statistical matching. Iet Computer Vision, 10(6), 551-558.

[5] Geiger, Martin Josef. (2017). A multi-threaded local search algorithm and computer implementation for the multi-mode, resource-constrained multi-project scheduling problem. European Journal of Operational Research, 256(3), 729-741.

[6] Chen, Z. Wang, X. Yan, K, & Zheng, J. (2020). Deep multi-scale feature fusion for pancreas segmentation from ct images. International Journal of Computer Assisted Radiology and Surgery, 15(3), 415-423.

[7] Scholz, P. Spitler, L. G, Hessels, J. W. T, Chatterjee, S, Cordes, J. M, & Kaspi, V. M, et al. (2016). The repeating fast radio burst frb 121102: multi-wavelength observations and additional bursts. Astrophysical Journal, 833(2), 177.

[8] Korus, P. & Huang, J. (2016). Multi-scale fusion for improved localization of malicious tampering in digital images. IEEE Transactions on Image Processing, 25(3), 1-1.