Using Visualization to Improve Clustering Analysis on Heterogeneous Information Network

Wenbo Wang[†], Yuwei Li[†], Feng Wang[‡], Xiaopei Liu[†*], Youyi Zheng[§†*]
[†]ShanghaiTech University
[‡]Jilin University
[§]Zhejiang University
{wangwb, liyw, liuxp}@shanghaitech.edu.cn,{wangfeng12}@mails.jlu.edu.cn,{zyy}@cad.zju.edu.cn,

***Abstract*—The exploration and analysis of data mining methodologies is an important task for effective knowledge discovery, especially in today's heterogeneous information networks. Previously presented approaches for mining optimization aim primarily at the improvements of time complexity, space complexity, accuracy, and robustness. We extend the state-of-the-art method by concentrating on user-availability and algorithm understandability. Specifically, we use Rankclus, a classic clustering algorithm as an example. After uncovering the unseen computing processes to be displayed in a visual form, the whole clustering processes are transparent to the users, which may help them more clearly and quickly understand how the algorithms are computed, how does each object influence one another. In addition, we use a density approach to intuitively simplify the discovery of data patterns, and through the visualized results, users can adjust algorithm parameters with or without professional training. Finally, we use another two visual techniques to improve the visualization quality: a heatmap matrix designed for checking the similarities of objects which are in the same cluster, and a DOItree implemented to further analyze the accuracy of the algorithms.**

***Index Terms*—data mining, heterogeneous information networks, visualization, Rankclus**

I. Introduction

The use of heterogeneous information networks (HINs) [1] [2] [3], such as social networks and the Web, has drawn extremely wide attentions in recent years. Meanwhile, an increasing number of exciting discoveries and successful applications are developed. They are able to find rich information hidden in heterogeneous links between entities in various fields: computer science, physics and biology, etc.. Among these developments, clustering and ranking are two prominent analytical techniques toward a better understanding of information network. Unlike homogeneous network, the algorithm for HINs has higher requirements, because HINs is ubiquitous and it has typed nodes and links, which would lead to a more informative discovery. In 2009, Rankclus [4], an algorithm designed for the data exploration of HINs, has been proposed by Sun and Han. They proposed the idea of exploring rank distribution for each cluster to improve clustering results. This algorithm can provide a higher accuracy of classifying entities in today's complicated information networks. Based on their work, many alternatives had been proposed gradually, such as NetClu [5], PathSim [6], etc..

During the past years, Rankclus based classification algorithms are contributed to solve the challenges in HINs, and we summarized that most of these improvements were derived from the perspective of algorithm design principles and features. For example. they consider the improvement of time complexity, space complexity, accuracy, and robustness. There are also some researchers doing works from the perspective of data volume and structures. Although they all make contributions in the development of information discovery in HINs, we find another two important elements are ignored: User practicality and algorithm understandability. User practicality represents whether the algorithm can be understood, and wheter it is used properly by the group of people without professional background; algorithm understandability represents whehter the algorithm can be read in an easy and clear way, rather than reading codes only. Based on this basic idea, we develop an idea that even a people new to the algorithms can grasp the computing processes quickly. Then, they can make contributions on the existing works without wasting too much efforts on the analysis. To fill this development gap, we propose using representative ways to make data mining processes to be transparent. The algorithms can be easily understood and adjusted, and hidden meaningful patterns can also be discovered. A well-recognized useful representative approach is visualization.

Generally, visualization techniques [7] [8] represent abstract data to reinforce human cognition. They enable the generation, interpretation, and manipulation of information through spatial representations. In other words, visualization can be an aid for researchers to better and more quickly understand the complicated computing processes, and they can also explore information knowledge. Therefore, to optimize algorithm performance, visualization is a desirable choice from the user's perspective. In our work, we dynamically monitor the computing processes and utilize a riverstream metaphor [9] [10] [11] [12], with variable-width trends, to explain the

[*]Corresponding author.

computing steps of clustering and ranking. To better understand the meaning of various river streams and clearly discover their transformations, we introduce a density approach to keep the continuity and clarity of each stream. In addtion, our visualization can help algorithm designer find the suitable parameters value more quickly and easily. For example, they can find when the algorithm is not having jittering problem, and what are the suitable vaules at this moment. Besides, we adopted another two visualization techniques, Heatmap Matrix [13] and DOItrees [14], to strengthen the analytical ability. These two techniques will cooperate with the main streams to discover hidden information patterns and relationships.

The main contributions of this work are:

- Using visualization to make the mining processes to be transparent, which will satisfactorily help users to better understand the process and adjust the algorithm parameters.
- Using density approach to connect neighbour objects that directly and visually solves the problem of information discontinuity.
- Using Jaccard Distance to get the similarity of identities that further understands the relationships between entities in Rankclus.
- Combining RiverStream, Heatmap Matrix and DOItree visualization techniques to cooperate with each other. Applying them to analyze the river patterns, which are generated during and after the whole computing processes. This could effectively present results overtime to users in an intuitive and manageable manner.

II. Related Works

This section reviews related works on optimization approaches of clustering algorithms and the application of visualization techniques in clustering in HINs.

Optimization Approaches of Clustering Algorithms. Clustering, as one of the most important questions of unsupervised learning, forms the basis for further knowledge mining. It has been used to group a set of objects, where the objects in a group are more similar to each other while differentiating with the objects in other groups. It has a wide acknowledgement in data mining, and has also been populaly applied and made contributions in various fields. Such as personalised environments, electronic commerce, and search engines, etc.. As a result, finding methods on how to improve their computing efficiencies becomes an important issue. Firstly, we take a look at how clustering algorithm had been improved in homogeneous information network, and then we consider its development in heterogeneous information network. Because networks that are homogeneous always containing the same type of objects and links. So the related clustering algorithms are developed quickly, which includes hierarchical clustering, centroid-based clustering, distribution-based clustering, and density based clustering. Take centroid-based clustering as an example, a classic method is K-means [15], and many algorithm alternatives had been proposed to cluster information based on this method. They are: KD-tree accelerated K-means [16]; speedup K-means with SVD decomposition [17]; the initialization clustering algorithm K-means++ [18] etc.. However, researchers found that homogeneous information netwrok usually extracts data from systems by ignoring the heterogeneity of objects and links, and sometimes only consider one type of relations among one type of objects. This would lead to the unaccuracy results. In order to improve the clustering quality, data can be modeled as heterogeneous information networks, which contain data with different types of objects and links, and the links and entities are all interconnected. In the heterogeneous information networks, traditional clustering algorithm could not get desirable classification results. But they can work together with ranking algorithms to get more accurate classification results.

The first proposed clustering algorithm in HINs is Rankclus, which is designed for bi-typed information network. Through using conditional ordering and mixed probability model, the algorithm can provide a clustering result with higher accuracy; followed by this, in order to deal with the data entity variety problem, Netclus [5] was proposed in 2013; then RankClass [19] algorithm was designed to let ranking and classification mutually enhance each other; and PathSim [6] was then discovered to find a better meta-path from many path choices. To summarize the existed optimization approaches on HIN clustering functions: One group of people are concentrated on data volume and structures; while the others are focused to optimize on the algorithm itself:(a) accuracy, which represents the algorithm can get the correct answers; (b) time complexity, which refers to the computational workload required to perform the algorithm; (c) robustness, which refers to the ability of an algorithm to respond to irrational data input and processing capabilities; (d) spatial complexity, which refers to the memory space the algorithm needs to consume; (e) generalization ability, which refers to the ability of machine learning algorithms to adapt the new samples. However, they all did not consider the possibility of improving optimization through usability, which contains user readability and algorithm understandability. In other words, if the algorithm can be read more intuitively, and the computing steps are more understandable. These might save users' time and efforts to further analyze and optimize algorithms, which can also extend the working oppotunities for more people, especially to the new in this area. Based on the above possibilities, in this paper, we propose the idea of using visualization to explain Rankclus algorithm. Through our experiments, we certify that visualization is effective on helping user understand algorithms.

Visualization techniques in HINs. A major contribution of visualization is using simple graphs to help human brain process complicated information. It is the explanation of data in a pictorial format rather than poring over in the number format. Through the prior researches on visualization, it has been certified that visualization effectively enables users to understand analysis process. Users can also grasp difficult concepts and identify information patterns through visualization results. During the past years, visualization techniques had been widely applied in many research fields. Specifically, in heterogeneous information networks, visualization has been effectively used in building data mining models. Rather than seeing the model as a black box, visualization transfers model outputs into meaningful graphical results and allows the user to interact with the results. We use two publicly recognized visualization techniques to explain this idea: One is drill-through [20]; the other is linking and brushing [21]. Drill-through can reveal additional details in a sub-model; while linking and brushing is able to highlight brushed data items in different representations. These two visualization techniques greatly help users understand how the model relates to original data, how the external contexts of the model are discovered, and how the validation are enhanced. On the other hand, in HINs, visualization is applied for model comparison. For example, if algorithms are required to be compared on the results of standard methods, such as computing time complexity, stability or computation size. Visualization techniques, bar charts or pie charts with other visual meaphors, such as colors and shapes [22] [23] [24] , are efficient to accomplish these tasks.

To differentiate the motivation of our work in using visualization in HINs, we proposed a novel approach to extend the applicability of visualization. It is using visualization to explain algorithms and making the processing details being transparent to users. In addition, through all results that are generated duing the processing processes, users can also discover information patterns and do further analysis on either algorithms or datasets. Specifically, to compare our idea with the existed approaches, we list three novities of our approach: Firstly, we propose a visualization based model to understand and analyze Rankclus algorithm; secondly, we concern that the uncertainty data may have a negative influence on the results, that they might cause discontinuity problems and then affect the pattern discovery in the analysis process.So we add an density approach to address this issue. This method is able to maintain visual effects without losing information pattern; thirdly, rankclus itself could not provide ranking results in each cluster. So according to the fundamental requirements of users, we combine Jaccard Similarity approach and Heatmap Matrix to further understand the inner relationships.

III. New Model in Algorithm Optimization

Algorithm optimization is a technique in the field of computer science. It refers to improve the relevant performance of the algorithm. Such as time complexity, space complexity, correctness, and robustness, etc.. With the arrival of the era of big data, more problems are coming to challenge the analysts' working efficiency, so how to improve the optimization algorithms also becomes an essential task.

Most of researchers chose to deal with the shortcomings of algorithm itself, either do numerous complicated programming experiments or apply various mathematical theories on the original theory. However, both of the two aspects are only suitable for researchers who are expert in information technology knowledge and programming. Sometimes strong mathematics background is also necessary. These limitations might decrease the working efficiencies of developing new algorithms and discovering algorithm issues. Because there is a higher possibility for people who are not very experienced in this area to have novel ideas on the problems and solve the problems from a different aspect. Therefore, in this paper, we propose using information visualization to help people, who are interested in algorithms but without professional backgrounds, to quicker and clearer understand an algorithm. Here is the new model.

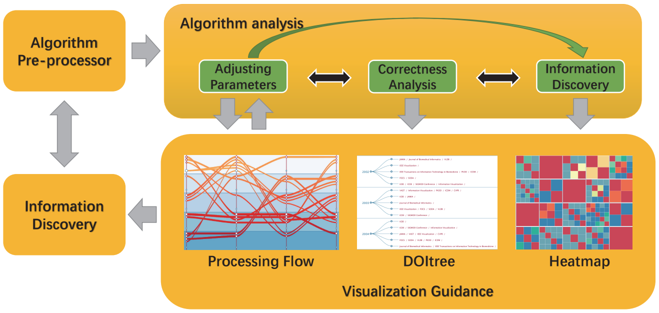

Fig. 1. New Model for Algorithm Analysis with Information Visualization

Fig.1 provides an overview of our model with its four main parts: Algorithm pre-processor, algorithm analysis, visualization guidance, and information discovery. The algorithm pre-processor is tasked to extract the basic idea of an algorithm, and then we can cooperate algorithm analysis and visualization guidance to complete main visual analysis. Specifically, in this part, we use three visualization techniques to explain the algorithm. First is the riverStream visualization (processing flow), which can discover how algorithms are running over objects, how parameters are affected over the algorithm, and how objects are affected with each other; Second is a DOITree visualization, which can verify the correctness of the algorithm. This is especially suitable for pretesting on a small part of dataset, while the classification results are alreay known. Then through DOITree visualization results, users can easily know whether the algorithm is running right or wrong; third is a Heatmap visualization, which is used for identifying inner relationships among objects per cluster. These three visualization techniques are working together to better represent the computing processes of an algorithm. We take Rankclus as an example: If we set the iteration number to 5, the riverstream will be drawed 5 times. And at any iteration step, we can check the result on DOITree. This will help users to understand how the results are changed at different iteration steps, and which step affects the results more. The last part of the model is information discovery. Users can interact with the results generated during computing processes to discover further information, such as data patterns or changing streams,etc..

IV. Rankclus Visualization

This section introduces how riverstream visualization helps to make the computing process of Rankclus to be transparent to users, and how algorithm performance are optimized through the dynamic observation of visualization results. The reminder of Section4 is structured as follows. Section A introduces Rankclus Algorithm. Section B discusses how does visualization guides user to understand and improve the clustering methods. Section C explains the further analysis of this visualizer.

A. Rankclus Algorithm

Network datasets had already transformed from homogeneous structure to heterogeneous structure since the data volume and variety are generated greater and quicker. A representative heterogeneous network can be seen as a bi-type directed graph [25] [26]. In this kind of graph, there are links among entities, and they are either having the same type or different types. To discover useful information from this kind of graph and better understand network properties, various analysis technologies are generated. Among these, ranking and clustering are two of the most important approaches. In this paper, we work on a fundamental algorithm Rankclus, which is also a classic clustering method in computer science bibliographic network.

Here is the basic principle of Rankclus [27] [28]. Suppose there are two entities, X and Y. The task is to ranking X_i. The network is randomly divided into K clusterings, and rank each of the cluster, then use the ranking result to be a metric in the following computing steps. After this, we use a mixture model to transfer each node into a K-Dimension vector, and classify it into X_k , which is the closest class with this node. Repeat the above steps, until the clustering results are stable. Specifically, when the algorithm are doing iterations, clustering results will be improved gradually. The similar objects(nodes) will get closer to each other. In addition, the more accurate of the clustering results are, the more correct the ranking results will be. These two functions are influenced each other. The followings are the main steps to implement Rankclus:

Step1: Ranking each cluster

In this step, there are two rules to follow to get better ranking results. The first rule is that if the author has a higher ranking, there will be a higher possibility for him/her to publish papers in the conferences with higher rankings; the second rule is that if a conference has a higher ranking, it might attract authors who are at higher ranking place. Specifically, these two rules illustrate that the value of each author is related with both the number of his/her publications and the weight of the conferences; the score of a conference depends both on the number of publications and the paper qualities; and the paper qualitis are related with the author ranking, that the higher of the author is ranking, the higher quality of the paper will be.

Mathematically, the principle can be explained by the following formulas: computing Y_i, then$\vec{r}_Y(j)$ can be computed by formula (1):

$$\vec{r}_Y(j) = \alpha \sum_{j=1}^{m} W_Y X(i,j) \cdot \vec{r}_X(i) + (1-\alpha) \cdot \sum_{j=1}^{m} W_Y Y(i,j) \cdot \vec{r}_Y(j) \tag{1}$$

where $\sum_{j=1}^{m} W_Y X(i,j)$ represents the number of the publications for the i_{th} author at the j_{th} conference; $\vec{r}_X(i)$ represents the score of the i_{th} conference; $\sum_{j=1}^{m} W_Y Y(i,j)$ represents the number of publications with the j_{th} coauthor; $\vec{r}_Y(j)$ represents the score of the j_{th} author.

Then normalize $\vec{r}_Y(j)$ by formula (2), denoted as $P_k(Y_j)$:

$$\vec{r}_Y(j) \leftarrow \frac{\vec{r}_Y(j)}{\sum_{j'=1}^{m} \vec{r}_Y(j')} \tag{2}$$

To be the same as the calucation with $\vec{r}_Y(j)$, $\vec{r}_X(i)$ can be calculated by formula (3) , and then it is normalized $\vec{r}_X(i)$ by fomula (4), which is denoted as $P_k(X_i)$:

$$\vec{r}_X(i) = \sum_{j=1}^{m} W_X Y(i,j) \cdot \vec{r}_Y(j) \tag{3}$$

$$\vec{r}_X(i) \leftarrow \frac{\vec{r}_X(i)}{\sum_{i'=1}^{m} \vec{r}_X(i')} \tag{4}$$

Step2: Clustering by mixture possibility model

Through step1, we rank each cluster,and get two conditional distributions for cluster K: one is $P_k(X_i)$, which is measured on the conference; the other is $P_k(Y_j)$; which is measured on the author. Then we use maximum likelihood estimation to estimate $P(z = k)$, refer to formula (5)-(7).

$$p(z = k|y_j, x_i, \theta) \propto p(x_i, yj|z = k)p(z = k|\theta^\circ) \tag{5}$$

$$p(z = k|y_j, x_i, \theta) = P_k(x_i)P_k(y_j)P^\circ(z = k) \tag{6}$$

$$p(z = k) = \frac{\sum_{i=1}^{m} \cdot \sum_{j=1}^{n} \cdot W_{XY}(i,j) p(z = k|x_i, y_j, \theta^\circ)}{\sum_{i=1}^{m} \cdot \sum_{j=1}^{n} \cdot W_{XY}(i,j)} \tag{7}$$

Then the algorithm calculates which cluster does the objects belongs to in each X, see formula(8).

$$p(z = k|x_i) = \frac{p_k(x_i)p(z = k)}{\sum_{r=1}^{k} p_q(x_i)p(z = q)} \tag{8}$$

Step3: Adjusting the clusters

In this step, the distance between X and the center of cluster K will be calculated, denoted as $d(x, X_k^(r))$. Cosine similarity is used for measuring the distance. See formula(9)-(10).

$$\vec{T}_X(k) = \frac{\sum_x \in x_k \cdot \vec{T}(x)}{|x_k|} \tag{9}$$

where $\vec{T}_X(k) = (p(z = 1|x_i), p(z = 2|x_i), ..., p(z = k|x_i)$ Then the distance d can be calculated by formula(11).

$$d(x, X_k) = 1 - cosine(x, X_k) \tag{10}$$

Repeat the three steps, until the algorithm is convergent.

B. *Visual Discovery of Rankclus*

The visualizer is able to transform complicated computing steps into comprehensive visualization results. In the visualization of Rankclus [29] [30] [31], a riverstream is presented to explain how the clustering results are generated, and how the parameters are controlled over each cluster. In the meanwhile, users can interact with the evolving trends to adjust the parameters, and dynamically changed data patterns can also be discovered during the visualizing process of each iteration. These stream-changing patterns can help users to understand more about dataset. In addition, we apply a DOItree to clearly verify whether the Rankclus is able to provide accurate results, and this technique always works together with the riverstream visualization. Finally, a Heatmap is designed to visually explain the inner correlations among objects in each cluster. The cooperation of these three visualization techniques makes algorithm optimization and information discovery much easier and quicker. We will describe them respectively.

a) Riverstream Visualization Panel: The RiverStream panel [10] visualizes the generating processes of the algorithm with a popular streamgraph metaphor, refer to Fig.2. It is an aesthetically pleasing and readily comprehensible visualization scheme, which is well established for visually integrating multiple time series. In addition, this metaphor makes it possible to link the changing among different iterations and cluster variations together, and visably discover their differences without breaking the visual effects. As a result, a comprehensible visualization is generated. Users can also interact with the interface naturally and smoothly. To explain more details of the algorithm: We imply a visual property, bubble shape, to represent the density classification for all objects. If the number of objects are more in a classification, the bubble size will be larger. Fig.3 gives a clear explanation of the relationships between bubble size and the number of objects. In this Figure, we use bubbles to represent the conferences. We can see that four conferences are chosen to be displayed, including IEEE Visualization, JAMIA, VAST and CVPR. They are all represented by streams with different shapes. It is obvious that IEEE Visualizaiton, JAMIA and VAST have the similar trends which are going up; while CVPR is going down at the same time period. And this phenomenon can also be found from the changing color of each stream. The color of IEEE Visualization, JAMIA and VAST are all changed to light organge; while CVPR is changed to dark orange. In the same figure, we can also see that the bubble size of four confernces are different, because their densities are different, they are 46.24

In addition, because the objects might have no value in a specific time period in a cluster, and the number of objects might be too much in another cluster. These can cause the results to be unclear. Therefore, we are facing two challenges: The visualization discontinuity and visual clutter. In our work, we propose using density approach to deal with these two issues without losing data accuracy. Specifically, density represents the percentage of each object in the whole objects.

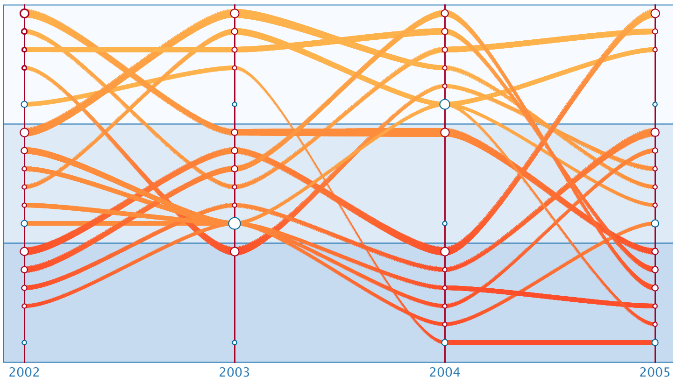

Fig. 2. An Example of RiverStream Visualization Panel

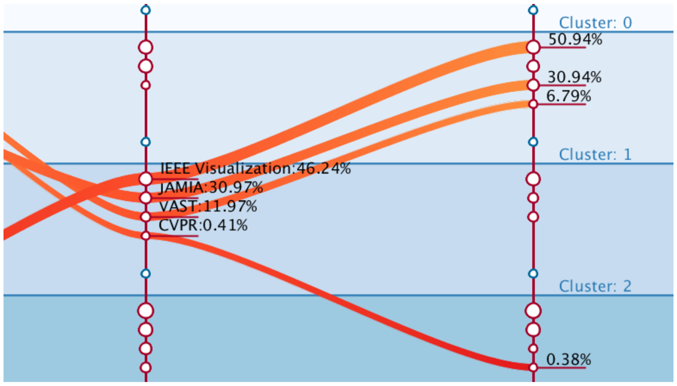

Fig. 3. An Example of Applying Bubble Visual Properties

We calculate the density of all objects, and category them by their values. If two objects have the same density, they will be gathered into one category. From the visualization point of view, our results only display bubbles that are the classifications of objects rather than displaying all objects. By using this approach, the number of river streams are decressed, and visual clutter problem is sloved. The effectiveness is especially obvious when the data volume is big. Besides, we set a special entity for each entity bubble cluster, the entity value is null, its color is highlighed as blue. This design is able to keep the continuity of each stream and remain a good visual affect. This also reconfirms the accuracy of objects classification results, refer to Fig.4.

Thirdly, we use interactive mechanism on riverstream to learn how objects(each stream represents an object) are transformed over time and what are the differences between them. In order to better illustrate the variation degree of each stream, we give each cluster a special colour. If an object is transferring to a different cluster, the stream colour will be gradually changed to the colour which is represented the destination cluster. This idea of drawing gradient colour can improve the visualization result, because users can more intuitively and transparently find their desired objects changing states. Fig.5 gives an illustratation of the interactive mechanism on riverstream. If the user click a stream, the color of the stream will be changed to green. In Fig.5, it is clear to observe that

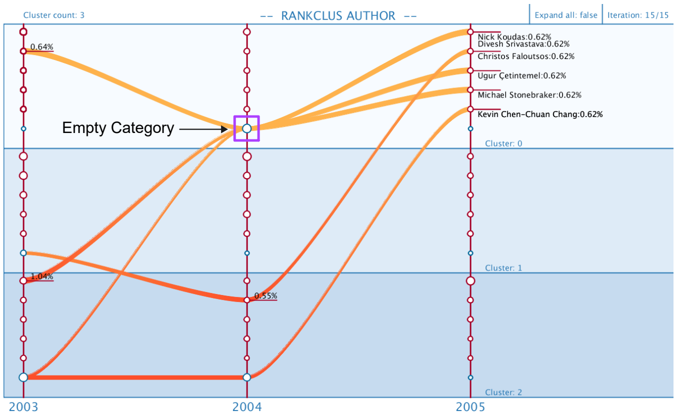

Fig. 4. An example of using null value to solve the visual discontinuity problem

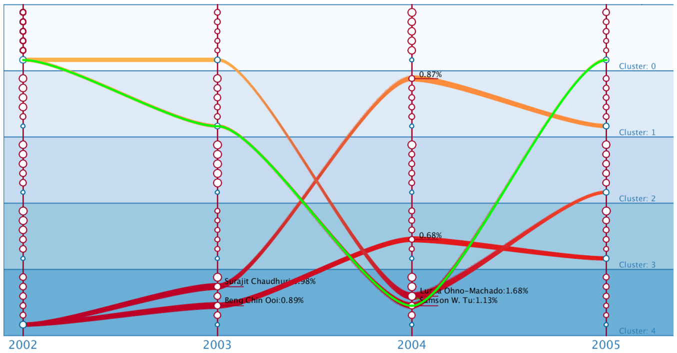

Fig. 5. An example of using color streams to differentiate the changing patterns of each cluster

this trend only has values in 2004. It is also interesting to find that although the values are all in a null cluster, they are distributed to different clusters. For example, in 2002, the null point belongs to cluster0; in 2003, it moves to cluster2; while in 2005, it goes back to cluster0. When this phenomenon apprears, it represents that this object could not be ranking at top k, which is the number of clusters. In Fig.5, we can also find that author Beng Chin Ooi and Surajit Chaudhuri are both transferred from cluster4 to cluster2, which means that cluster 4 and cluster 2 are more similar with each other, because it is less likely that researchers at the same time period change their research interests to a same research area.

b) DOITrees Visualization Panel: In reality, hierarchical structures provide ways to present complex structures in a simplified form. In the past decades, various 2D hierarchical structures visualization techniques have been proposed, such as Treemaps [32], Space-Optimized Tree [33], EncConTree [34], and SpaceTree [35] etc.. Among these, Degree-of Interest Trees (DOITrees) is one of the most popular techniques for large tree visualizations. It can provide simple and clear preview icons for the summarization of the complex structures. Therefore, in this paper, we choose DOITrees to display the different clustering results when the algorithm is doing calculatation. This means that each iteration has a DOITree result, and when the algorithm finished all calculation, the final result will also be displayed in a DOITrees form. This visualization technique mainly cooperates with riverstream to verify the correctness of an algorithm, refer to Fig.6.

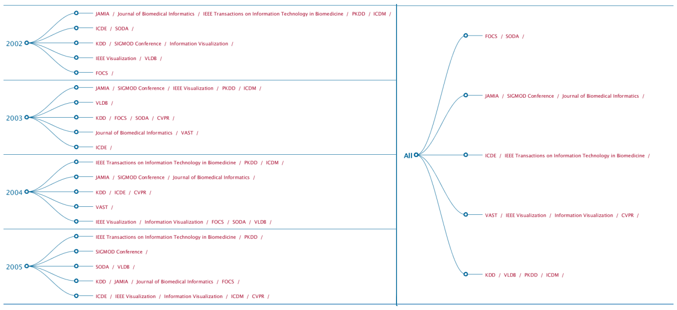
2002
JAMIA / Journal of Biomedical Informatics / IEEE Transactions on Information Technology in Biomedicine / PKDD / ICDM /
ICDE / SODA /
KDD / SIGMOD Conference / Information Visualization /
IEEE Visualization / VLDB /
FOCS /
2003
JAMIA / SIGMOD Conference / IEEE Visualization / PKDD / ICDM /
VLDB /
KDD / FOCS / SODA / CVPR /
Journal of Biomedical Informatics / VAST /
ICDE /
2004
IEEE Transactions on Information Technology in Biomedicine / PKDD / ICDM /
JAMIA / SIGMOD Conference / Journal of Biomedical Informatics /
KDD / ICDE / CVPR /
VAST /
IEEE Visualization / Information Visualization / FOCS / SODA / VLDB /
2005
IEEE Transactions on Information Technology in Biomedicine / PKDD /
SIGMOD Conference /
SODA / VLDB /
KDD / JAMIA / Journal of Biomedical Informatics / FOCS /
ICDE / IEEE Visualization / Information Visualization / ICDM / CVPR /
All
FOCS / SODA /
JAMIA / SIGMOD Conference / Journal of Biomedical Informatics /
ICDE / IEEE Transactions on Information Technology in Biomedicine /
VAST / IEEE Visualization / Information Visualization / CVPR /
KDD / VLDB / PKDD / ICDM /

Fig. 6. An example of DOITree Visualization - time based results

c) Heatmap Visualization Panel: It is important to understand the cluster processes, and to verify the correctness of the results through the cooperation of riverstream and DOITree visualizatioin. It is also essential to discover the relationships among objects which are in the same cluster. In this paper, we apply Jaccard Distance to measure the similarity distance. The main reason is that this similarity measurement only concentrates on whether the objects are similar on the same features, and this matches our motivation to find differences among the objects which are in the same cluster.

To better understand the similarity measurement result, we apply the idea of heatmap visualization technique. Refer to Fig.7. A heatmap is a graphical representation of data where the individual values contained in a matrix, and the differences of values are represented as various colours. In our paper, in order to explain all relationships among objects, we use a matrix to express the similarities. Each square represents the similarity of two objects. As for the colour schemes, considering human perceptual advantages and disadvantages, we use a principle of colour contrast: while the similarity value is larger, the colour will be brighter. The colours are chosen from light blue to dark red, see the legend in Fig.7.

d) The Cooperation of Three Visualization Techniques: Riverstream, DOITree and heatmap are working together to finish the task of unveiling the processes of algorithms and discovering information patterns. In our Rankclus visualization, Riverstream is the main visualizer to represent the computing process of Rankclus. User can choose attributes to be the main visualized objects, either author or conference. Suppose we choose conference as the target. In the beginning, user can set the iteration number and cluster number,then the visualizer will give the clustering results for each iteration in the stream format. During the whole computing process, the user can detect the changing frequencies of each cluster. Suppose since iteration 5, the riverstream could not changed anymore. Then, users can adjust the iteration value to decrease the possibility of wasting computing resources. As the changes in the riverstream might be small, and are not easy to find by our eyes. So we can use DOITree panel to display cluster

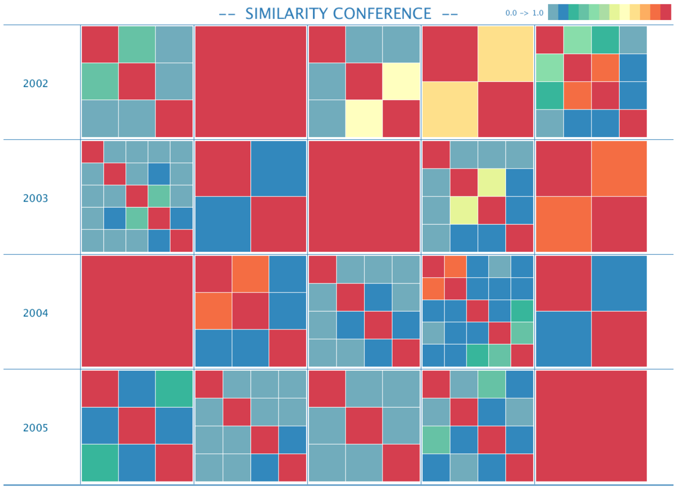

Fig. 7. An example of Heatmap Visualization Panel

TABLE I
ILLUSTRATION OF THE TEST DATASET

CATEGORY	CONFERENCE	ATTRIBUTES
DataMining	KDD PKDD ICDM	Conference Name Paper Detail Author Name Publishing Time (2002-2005)
DataBase	SIGMOD Conference VLDB ICDE	
Medical Informatics	Journal of Biomedical Informatics IEEE Transactions on Information Technology in Biomedicine JAMIA	
Theory	FOCs SODA	
Visualization	Information Visualization IEEE Visualization VAST CVPR	

results of each iteration, the results will be displayed in text, and we can make sure whether the cluster result will be affected if we change the iteration number. After we adjust the parameters for the algorithm, we are able to use heatmap and DOITree together to check the relationships among all entities, refer to Fig.6 and Fig.7.

C. Further Analysis of Rankclus on Aminer

We apply this visualizer on a network dataset Aminer (https://www.aminer.cn/data) to understand and optimize Rankclus algorithm. This dataset contains paper information, author information, paper citation, author collaboration, conference information, etc.. In the test, we chose a data subset which contains paper information, author information and research fields, the test dataset details are illustrated in TABLE.I.

First, we disorder the dataset to make sure that the conferences are not in the same cluster. Then run the Rankclus algorithm to get clustering results. A stream can represent either author or conference, which depends on the user requirements. Through the changes of riverstreams, users can discover how conferences are transformed among clusters, and how authors are transformed among research fields. In addition, users can check the results in DOItree and then compare the results with the correct cluster result in TABLE.I. If the result is not correct, the users can adjust the value of alpha and the iteration number until the result is accurate. In this step, the users can visually understand how algorithm parameters (alpha and iteration number) are affected the algorithm; what are the suitable values for parameters to make algorithm to be stable, which is with less computing cost, and without losing information accuracy. TABLE.II is the test results for a Rankclus algorithm on Aminer datasets. From the previous researches on Rankclus algorithm, it has been discovered that the range of alpha shoule be from 0.95 to 1.00, and a suitable iteration number would be around 20. And through our test, a better visualization results can be received when alpha equals to 0.98, and the iteration value is 20, which is the same conclusion with the Rankclus researchers.

TABLE II
ANALYSIS of ALGORITHM PARAMETERS

Alpha	Iteration times	Clustering Accuracy
0.99	15	Theory
	18,19	Database
	20,21,22	-
0.98	15,22	-
	18	Medical Informatics
	19	Visualization
	20	Database Visualization
	21	Medical Informatics Visualization
0.96	15,22	-
	18,19,21	Theory
	20	Database Medical Informatics
0.95	15,19,20	Theory
	18	-
	21	Database
	22	Medical Informatics

V. Conclusions

In this paper, we discussed the developments and challenges about algorithm optimization in HINs. In association with these challenges, we discuss the methodologies and techniques proposed in recent years. Compared with previous solutions, we have presented a visual analysis technique to help users understand and optimize algorithms. Specifically, our approach is unique in two aspects. First, it allows user, who are not professional in algorithms, are able to better and more quickly understand how algorithms are working on datasets; second, it enables users to interact with algorithms. This method supports an interactive analysis cycle with the cooperation of three visualization techniques. And through the case study on Aminer, we have demonstrated the usability of our idea in visually analysing algorithm computing processes.

Our design also has some limitations on the analysis of Rankclus. First of all, it is not clear to display how the attributes are affected each other. Take Aminer as an example, user can see the processes of author ranking and conference ranking, but cannot clearly discover how author and conference are affected each other at the same time. As a consequence, users have to explore visual results thoroughly to find patterns. Second, it is still not easy to discover stream patterns when the dataset is extremely large in our current system.

Based on our idea of using visualization to guide the development of algorithm. In the future, on the one hand, we will improve the visualization of existing Rankclus algorithm; on the other hand, we are interested in discovering suitable visualization techniques to explain Neural network algorithm [36], and make a transparent way to see how neural networks learn.

Acknowledgment

We thank all reviewers for their valuable comments. This work was supported in part by the National Natural Science Foundation of China NO. 61502306, the China Young 1000 Talents Program.

References

[1] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S Yu. Mining heterogeneous information networks. In *Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10), Washington, DC*, 2010.

[2] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.

[3] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: a structural analysis approach. *Acm Sigkdd Explorations Newsletter*, 14(2):20–28, 2013.

[4] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[5] Elena Baralis, Andrea Bianco, Tania Cerquitelli, Luca Chiaraviglio, and Marco Mellia. Netcluster: A clustering-based framework to analyze internet passive measurements data. *Computer Networks*, 57(17):3300–3315, 2013.

[6] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.

[7] Usama M Fayyad, Andreas Wierse, and Georges G Grinstein. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.

[8] Mike Cammarano, Xin Dong, Bryan Chan, Jeff Klingner, Justin Talbot, Alon Halevey, and Pat Hanrahan. Visualization of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1200–1207, 2007.

[9] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE transactions on visualization and computer graphics*, 8(1):9–20, 2002.

[10] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE transactions on visualization and computer graphics*, 22(1):190–199, 2016.

[11] Dongning Luo, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE transactions on visualization and computer graphics*, 18(1):93–105, 2012.

[12] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1763–1772, 2014.

[13] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[14] Quang Vinh Nguyen, Simeon Simoff, and Mao Lin Huang. Using visual cues on doitree for visualizing large hierarchical data. In *Information Visualisation (IV), 2014 18th International Conference on*, pages 1–6. IEEE, 2014.

[15] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[16] Dan Pelleg and Andrew Moore. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–281. ACM, 1999.

[17] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[18] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7):622–633, 2012.

[19] Ming Ji, Jiawei Han, and Marina Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1298–1306. ACM, 2011.

[20] James Ahrens, Kristi Brislawn, Ken Martin, Berk Geveci, C Charles Law, and Michael Papka. Large-scale data visualization using parallel data streaming. *IEEE Computer graphics and Applications*, 21(4):34–41, 2001.

[21] Daniel A Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[22] Nathan Kogan, Kathleen Connor, Augusta Gross, and Donald Fava. Understanding visual metaphor: Developmental and individual differences. *Monographs of the Society for Research in Child Development*, pages 1–78, 1980.

[23] Peter R Keller, Mary M Keller, Scott Markel, A John Mallinckrodt, and Susan McKay. Visual cues: practical data visualization. *Computers in Physics*, 8(3):297–298, 1994.

[24] Hermine Feinstein. Meaning and visual metaphor. *Studies in Art Education*, 23(2):45–55, 1982.

[25] Yizhou Sun and Jiawei Han. Integrating clustering with ranking in heterogeneous information networks analysis. In *Link Mining: Models, Algorithms, and Applications*, pages 439–473. Springer, 2010.

[26] Dipak R Pardhi and Akhilesh A Waoo. An efficient ranking based clustering algorithm. *International Journal of Engineering and Advanced Technology (IJEAT)*, 1(1), 2011.

[27] Xing Le. Rankclus on directed graph and its application. *China's Outstanding Master's Degree thesis*, 7, 2013.

[28] Huajie Shao, Jinda Han, and Sida Li. Highsim: Highly effective similarity measurement in large heterogeneous information networks.

[29] Yintao Yu. Ivis: Search and visualization on heterogeneous information networks. 2011.

[30] TAO Jianwen. Rchig: an effective clustering algorithm with ranking. *Journal of Software*, 4(4), 2009.

[31] Zhiguo Zhu, Jingqin Su, and Liping Kong. Measuring influence in online social network based on the user-content bipartite graph. *Computers in Human Behavior*, 52:184–189, 2015.

[32] Ben Shneiderman and Martin Wattenberg. Ordered treemap layouts. In *Information Visualization, 2001. INFOVIS 2001. IEEE Symposium on*, pages 73–78. IEEE, 2001.

[33] Quang Vinh Nguyen and Mao Lin Huang. A space-optimized tree visualization. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 85–92. IEEE, 2002.

[34] Mao Lin Huang, Quang Vinh Nguyen, Wei Lai, and Xiaodi Huang. Three-dimensional enccon tree. In *Computer Graphics, Imaging and Visualisation, 2007. CGIV'07*, pages 429–433. IEEE, 2007.

[35] Catherine Plaisant, Jesse Grosjean, and Benjamin B Bederson. Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *The Craft of Information Visualization*, pages 287–294. Elsevier, 2003.

[36] Martin T Hagan, Howard B Demuth, Mark H Beale, et al. *Neural network design*, volume 20. Pws Pub. Boston, 1996.