SCIENCE CHINA
Information Sciences

• MOOP •

August 2020, Vol. 63 184101:1–184101:3
https://doi.org/10.1007/s11432-019-2647-3

Ordered matrix representation supporting the visual analysis of associated data

Yi CHEN[1*], Cheng LV[1], Yue LI[1], Wei CHEN[2] & Kwan-Liu MA[3]

[1] *Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing* 100048*, China;*
[2] *State Key Laboratory of CAD&CG, Zhejiang University, Zhejiang* 310058*, China;*
[3] *Department of Computer Science, University of California, Davis* 95616*, USA*

Received 14 June 2019/Revised 3 September 2019/Accepted 9 September 2019/Published online 26 March 2020

Citation Chen Y, Lv C, Li Y, et al. Ordered matrix representation supporting the visual analysis of associated data. Sci China Inf Sci, 2020, 63(8): 184101, https://doi.org/10.1007/s11432-019-2647-3

Associated data, which refer to sets of entities with specific relations and relational weights that can usually be expressed in a relational matrix, are found in many fields. Two typical examples are the pesticide residue dataset in food-safety, in which pesticides are associated with agricultural products [1], and E-transaction dataset, in which there is special relation between buyer and seller [2]. However, on large data scales, finding the key entities or mine hidden patterns in such data is difficult and time consuming. An ordered matrix can help analysts quickly locate the entity of interest. When ranking the entities, the relations and their weights should both be considered. However, the existing ranking algorithms such as PageRank [3], consider only the relations while ignoring their weights. In this paper, we propose a ranking algorithm called RW-Rank, in which RW stands for both relations and weights. RW-Rank is inspired by the PageRank algorithm [3] and is available for create ordered relational matrices. Emergent visual analysis technology can improve the efficiency of complex associated-data analysis [4]. Using this new technology, we design and implement a visual analysis system called Rank-Vis for analysing associated data by the RW-Rank algorithm.

Data model. We abstract the associated data as a bipartite graph, which divides all vertices into two independent sets. Every edge connects a vertex in one set to a vertex in the other set [5]. The bipartite graph is defined as $G = (P, R, W)$, in which $P = \{P_1, P_2, \ldots, P_i, \ldots, P_m\}$ is one vertex set, $R = \{R_1, R_2, \ldots, R_j, \ldots, R_n\}$ is the other vertex set, and $W = \{w_{11}, w_{12}, \ldots, w_{ij}, \ldots, w_{mn}\}$ is the set of edges. The matrix representation is given by (1). Here, $P_i = \{w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in}\}$ represents an entity in P, i.e., a row vector, and $R_j = \{w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{mj}\}$ represents an entity in R, i.e., a column vector. P_i and R_j are weighted and related through w_{ij}. If $w_{ij} = 0$, then no relation exists between the two entities.

$$G = \begin{array}{c} \\ P_1 \\ P_2 \\ \vdots \\ P_i \\ \vdots \\ P_m \end{array} \begin{array}{c} \begin{array}{cccccc} R_1 & R_2 & \cdots & R_j & \cdots & R_n \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & \cdots & \cdots & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & \cdots & \cdots & w_{2n} \\ \ldots & \ldots & \ddots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & w_{ij} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ddots & \ldots \\ w_{m1} & w_{m2} & \cdots & \cdots & \cdots & w_{mn} \end{bmatrix} \end{array}. \quad (1)$$

RW-Rank algorithm. We define the RW value of an entity as an index, indicating the importance of this entity in the relational structure. The RW is calculated by (2), where RW_i is the RW value of entity P_i and RW_j is the RW value of entity R_j. O_j is the number of non-zero elements (i.e., the out-degree of vertex R_j). Similarly, O_i is the out-degree of vertex P_i.

[*] Corresponding author (email: chenyi@th.btbu.edu.cn)

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2020

$$\mathrm{RW}_i = \sum_{j=1}^{n} w_{ij} \frac{\mathrm{RW}_j}{O_j}, \ \mathrm{RW}_j = \sum_{i=1}^{m} w_{ij} \frac{\mathrm{RW}_i}{O_i}. \quad (2)$$

The RW-Rank algorithm ranks the row and column vectors in the matrix according to their RW values. We define two sets, A and B, that record the RW values of all row and column vectors in the relational matrix G, respectively. Specifically, $A = \{a_1, a_2, \dots, a_i, \dots, a_m\}$ and $B = \{b_1, b_2, \dots, b_j, \dots, b_n\}$, where a_i and b_j are the RW values of row vector P_i and column vector R_j, respectively. The RW-Rank algorithm is proceeded as follows.

- **Step 1.** Construct the relational matrix G and set the threshold θ, at which iterations are ended. The threshold is set by users.
- **Step 2.** Initialize the RW values of all row and column vectors to 1, i.e., let $a_i = 1$ $(i = 1, \dots, m), b_j = 1$ $(j = 1, \dots, n)$. Initialize θ to a small value, such as 0.001.
- **Step 3.** Update the RW values of all row and column vectors. Compute their elements by (3) and (4), respectively, and insert them as $A* = \{a_1^*, a_2^*, \dots, a_i^*, \dots, a_m^*\}$ and $B^* = \{b_1^*, b_2^*, \dots, b_j^*, \dots, b_n^*\}$, respectively.
- **Step 4.** Calculate the difference ε in the RW values before and after the update by (5).
- **Step 5.** If $\varepsilon > \theta$, then let $A = A^*$ and $B = B^*$, jump to Step 3 and continue the iterations; otherwise, end the iteration and execute Step 6.
- **Step 6.** Rank the row and column vectors of the relational matrix G according to their RW values in A^* and B^*, respectively. Finally, create an ordered relational matrix.

$$a_i^* = \mathrm{RW}_i = \sum_{j=1}^{n} w_{ij} \frac{\mathrm{RW}_j}{O_j} = \sum_{j=1}^{n} w_{ij} \frac{b_j}{O_j}, \quad (3)$$

$$b_j^* = \mathrm{RW}_j = \sum_{i=1}^{m} w_{ij} \frac{\mathrm{RW}_i}{O_i} = \sum_{i=1}^{m} w_{ij} \frac{a_i}{O_i}, \quad (4)$$

$$\begin{aligned}\varepsilon &= \|A^* - A\| + \|B^* - B\| \\ &= \sum_{i=1}^{m} (a_i^* - a_i) + \sum_{j=1}^{n} (b_j^* - b_j). \quad (5)\end{aligned}$$

Rank-Vis system. Based on the RW-Rank algorithm, Rank-Vis assists users to locate the key entities and analyse the relations between them. We verify the effectiveness of this approach on two case studies: a pesticide residue dataset for introducing the system in detail, and a students' course achievement dataset. As shown in Figure 1, the user interface provides coordinated multiple views showing different aspects of the residue data: ordered matrix heatmaps, word clouds, pie charts, and parallel coordinates. Interaction techniques, such as filter, highlight and lasso selection, are also provided for analysis.

In the parameter panel (E), users can select a dataset through "DataSet" tag. After selecting the pesticide residue dataset, users can filter the data of interest by selecting city, category, and year. For the present case study, we choose the vegetable data of City A in 2014, which includes 12 agricultural products and 48 pesticides.

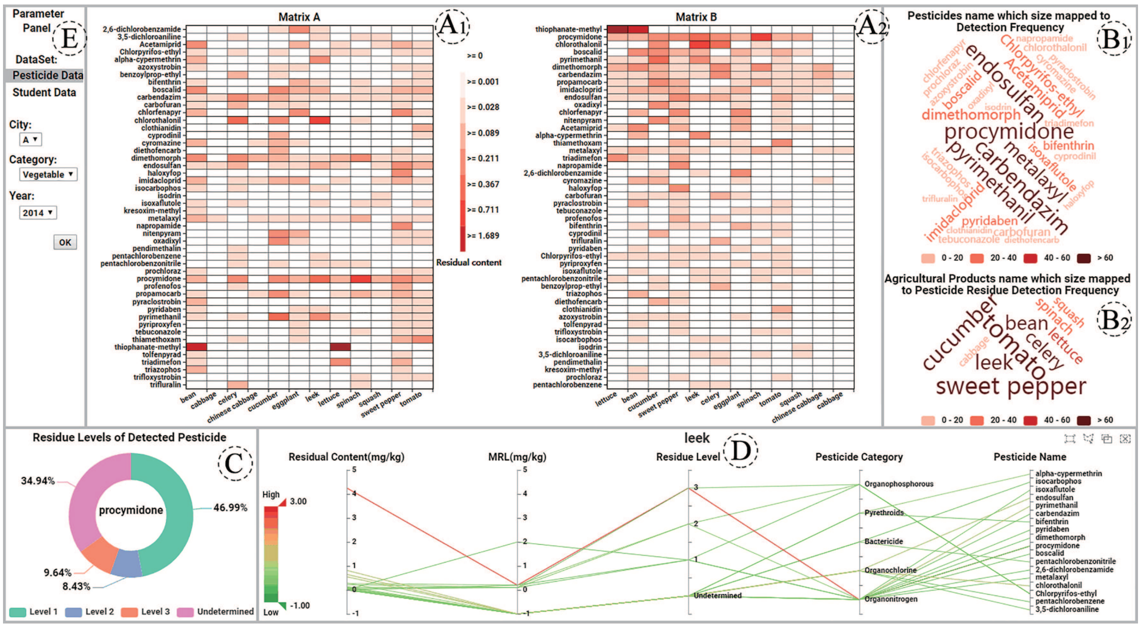
The matrix heatmap (A_1, A_2) shows the residual content and relations between the pesticides and agricultural products. Each row and column vector represents a pesticide entity and an agricultural product entity, respectively. The darker the colour, the higher is the residual content. In matrix A, users can quickly locate the names of pesticides and agricultural products, which are listed in alphabetical order. Meanwhile, the rows and columns in matrix B are ordered by their RW values. The cell colour gradually lightens from the upper left corner to the lower right corner. Users can quickly locate pesticides with high residual contents and seriously pollute agricultural products. When users click a cell in matrix A, the corresponding cell in matrix B is highlighted, and vice versa. In this case study, the residual content of thiophanate-methyl in lettuce is the highest among the entities in matrix A, reaching 2.446 mg/kg (as can be seen by hovering the mouse over this entity). In matrix B, thiophanate-methyl and procymidone (in rows 1 and 2, respectively) received the highest RW values. The high RW reflects a high residual content in thiophanate-methyl, and a high detection frequency in procymidone. Lettuce (in column 1) receives the highest RW among the agricultural produces, denoting heavy pollution of this product.

The word clouds (B_1 and B_2) display the names of all row and column vectors, i.e., the names of all pesticide and agricultural products, respectively. The detection frequency of pesticide is highest for procymidone (83 occasions), and that of the agricultural product is highest for tomatoes (143 occasions).

The pie chart (C) shows the percentages of the four residue levels, defined as follows (where RC and MRL denote the residual content, and the maximum residue limit, respectively):

- Level 1: RC $<$ 0.1MRL;
- Level 2: 0.1MRL $\leqslant$ RC $<$ MRL;
- Level 3: RC $\geqslant$ MRL;
- Undetermined: The MRL is unspecified.

The levels 1, 2, 3 and undetermined residue levels of procymidone are 46.99%, 8.43%, 9.64%, and 34.94%, respectively.



Parameter Panel

DataSet: Pesticide Data

City: A

Category: Vegetable

Year: 2014

Figure 1 (Color online) Screenshot of the Rank-Vis system for visually analysing the pesticide-residue dataset selected in the parameter panel (E). (A_1) The row and column vectors in matrix A are ranked in alphabetical order. (A_2) Matrix B is the same matrix with the row and column vectors reordered by the proposed RW-Rank algorithm. (B_1) and (B_2) are the word clouds, in which the sizes of the pesticide and agricultural product names are mapped to the detection and pesticide residue detection frequencies, respectively. (C) Pie chart showing the percentages of four residue levels (levels 1, 2, 3 and undetermined) in different agricultural products selected by clicking their names in (B_1). (D) Parallel coordinate shows the detailed information of a specific agricultural product selected by clicking its name in (B_2).

The parallel coordinate (D) shows the residual content, MRL, the residue level, pesticide category and pesticide name in the agricultural product selected by clicking its name in B_2. Along the MRL axis, −1 indicates an undetermined MRL. The data range can be selected using the colour spectrum and the lasso tool.

Exploring and analysing the above dataset by the Rank-Vis system, we find that: (1) thiophanate-methyl and procymidone are the critical pesticides; (2) lettuce and beans are the most seriously polluted agricultural products; (3) carbendazim is detected in all agricultural products; and (4) the MRLs of some pesticides have not yet been formulated.

The main contributions of this article are summarized below.

- We presented our RW-Rank algorithm, which creates an ordered relational matrix of relations and their weights.
- We defined the RW-value for quantifying the importance of an entity, and presented its calculation method.
- We designed and implemented a visual analysis system for associated data analysis. The system, called Rank-Vis, can help analysts to quickly locate the important entities.

The students' course achievement dataset, is analysed and discussed in the supplemental documents. The proposed method can analyse the data associations in diverse fields.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2018YFC1603602), National Natural Science Foundation of China (Grant Nos. 61972010, 61772456, 61761136020), and Basic Research Project of the Ministry of Science and Technology (Grant No. 2015FY111200).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen Y, Du X M, Yuan X R. Ordered small multiple treemaps for visualizing time-varying hierarchical pesticide residue data. Vis Comput, 2017, 33: 1073–1084
2 Xie C, Chen W, Huang X X, et al. VAET: a visual analytics approach for E-transactions time-series. IEEE Trans Visual Comput Graph, 2014, 20: 1743–1752
3 Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab
4 Mei H H, Ma Y X, Wei Y T, et al. The design space of construction tools for information visualization: a survey. J Visual Lang Comput, 2018, 44: 120–132
5 Du X M, Chen Y, Li Y. TransGraph: a transformation-based graph for analyzing relations in data set. J Comput-Aided Des Comput Graph, 2018, 30: 79–89