2010 Second International Workshop on Education Technology and Computer Science

Predicting Internet Network Distance using ISOMAP

Liu Xianglong, Lou Yihua, Liang Yuan, Shan Baosong
State Key Laboratory of Software Development Environment
Beihang University
Beijing, P.R.China
{xlong_liu, louyh, liangyuan, shanbs}@nlsde.buaa.edu.cn

***Abstract*—Since coordinate-based methods for network distance prediction can estimate distances more accurately and effectively than previously proposed methods, they have been widely studied and used in Internet applications. However, there still exist at least three problems unsolved: to find a embedding low-dimensional Euclidean space best preserving distance information, to determine the dimension of embedded Euclidean space, and to reduce time and parametric complexity derived from iterative optimizing process. This paper proposes a new coordinate-based method using ISOMAP to address these problems. ISOMAP estimates distances between nodes by their shortest path distance and employs Multidimensional Scaling (MDS) which uses matrix decomposition to find nodes' coordinates in embedding Euclidean space best preserving distances. MDS avoids the complexity of optimization and helps exploit the dimension size of embedding space according to information preservation. Discussion and experiments have proved that the proposed method performs faster and more accurately than the Global Network Positioning (GNP) does.**

***Keywords*—network coordinates; distance prediction; shortest path distance; ISOMAP; multidimensional scaling**

I. Introduction

Since the performance of distributed Internet applications can be improved significantly if the network distances among nodes can be estimated accurately, there has been much concentration on this area in the literature. Most of these previous research [1][2][3][4][5] for distance estimation can be classified into coordinate-based methods, of which the Global Network Positioning (GNP) [2] pioneers Internet coordinate distance estimation which uses a set of fixed landmark nodes as the reference nodes that can be probed by nodes joining the system afterwards. Lots of experiments have shown that GNP performs well on the real-world data. Many works follow GNP. For example, Vivaldi [3] uses a different method stabilizing spring systems instead of optimizing error minimization to obtain the optimal coordinates in the embedded space, while on the basis of GNP some improve scalability, accuracy, and other aspects to certain extent. Because coordinate-based methods can predict distances more accurately and effectively than previously proposed methods do, they benefit applications including nearest server selection, data distribution, etc.

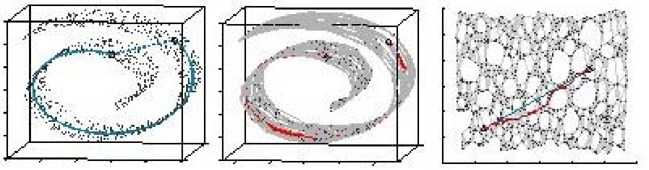

Fig. 1. The Swiss roll data set

However, there are still at least three problems remaining unsolved: (1) Mapping from network distance into low-dimensional Euclidean space. Intuitively Internet nodes might distribute in a sphere space or other non-linear space, therefore for cases shown in Fig. 1 [6], the predicted Euclidean distances (length of dashed line) using distance function in low dimensional Euclidean space may not accurately reflect the measured distances (length of solid curve). (2) Determining the dimension of embedded Euclidean space. Experimental results imply that network distances can be modeled by 5-7 dimensional Euclidean space [5], but GNP and other methods determine the size experimentally which means that they are not practicable and adaptive to different data. (3) Reducing time and parametric complexity. GNP's iterative optimizing process is time-consuming and needs parametric configuration to promise fast convergency and optimal solution.

To solve these problems, in this paper we propose a novel method for network distance prediction based on ISOMAP [6]. ISOMAP estimates the distance between two nodes by their geodesic distance (i.e. the length of shortest path), which appears more close to the measured one in the network distance space [8]. Then Multidimensional Scaling (MDS) [7] of low complexity is employed to find nodes' coordinates best preserving distances and to exploit the dimension size of embedding Euclidean space according to distance preservation. Our method consists of two parts similar to GNP: (1) in the first part landmarks' coordinates of specified dimension determined by MDS are obtained, and the optimizing complexity in GNP can be avoided by matrix decomposition; (2) in the second part coordinates of a new node joining can be found by optimizing the error function.

The rest of the paper is organized as follows: An overview of the GNP framework is presented in Section 2. In Section 3

Supported by a grant from the Major State Basic Research Development Program of China (973 Program) (No.2005CB321903) and a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2007AA010301-02)

978-0-7695-3987-4/10 $26.00 © 2010 IEEE
DOI 10.1109/ETCS.2010.245

a new coordinates-based method using ISOMAP for distance prediction is described in detail. Section 4 evaluates the proposed method with respect to different measurements and discusses the results. Finally, we conclude in section 5.

II. Global Network Positioning

The Global Network Positioning (GNP) is an approach proposed to estimate the Internet network distances based on absolute coordinates. The key idea of GNP is to model the Internet as a geometric space (e.g. a 3-dimensional Euclidean space) and to map Internet nodes into points in this space.

GNP comprises of two parts. In the first part, a small set of nodes $(\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N)$ named landmarks are chosen and their coordinates in a specified space are obtained by optimizing the following objective function:

$$\min \sum_{\mathcal{L}_i, \mathcal{L}_j \in \{\mathcal{L}_1, .., \mathcal{L}_N\} | i>j} \epsilon^2(d_{\mathcal{L}_i\mathcal{L}_j}, \hat{d}_{\mathcal{L}_i\mathcal{L}_j}) \qquad (1)$$

Where $d_{\mathcal{L}_i\mathcal{L}_j}$ and $\hat{d}_{\mathcal{L}_i\mathcal{L}_j}$ are measured distance (round-trip times) and predicted distance between $\mathcal{L}_i$ and $\mathcal{L}_j$ respectively, and $\epsilon(\cdot)$ represents the error function. For example, GNP uses relative error function $\epsilon(d_{\mathcal{L}_i\mathcal{L}_j}, \hat{d}_{\mathcal{L}_i\mathcal{L}_j}) = \frac{d_{\mathcal{L}_i\mathcal{L}_j} - \hat{d}_{\mathcal{L}_i\mathcal{L}_j}}{d_{\mathcal{L}_i\mathcal{L}_j}}$.

In the second part of GNP framework, these landmarks serve as a set of reference coordinates to which distances can be measured for each non-landmark node $\mathcal{H}$. Following this the coordinates of the node can be obtained by minimizing the following objective function:

$$\min \sum_{\mathcal{L}_i \in \{\mathcal{L}_1, .., \mathcal{L}_N\}} \epsilon^2(d_{\mathcal{H}\mathcal{L}_i}, \hat{d}_{\mathcal{H}\mathcal{L}_i}) \qquad (2)$$

Where $\epsilon(\cdot)$ is the same error function.

After coordinates of all the nodes are computed, we can predict distances between any two nodes which benefits a variety types of network applications.

III. Distance Prediction Using ISOMAP

Since experiments in [2][3][4][5] have shown that Euclidean space for network distances is reasonably accurate for dataset they studied, what our concentration turns to is how to construct a low-dimensional Euclidean space best preserving network distances. It can be viewed as an embedding problem. Reference [2] calculates nodes' coordinates using non-metric Multidimensional Scaling (MDS), while reference [5] employs Principle Component Analysis (PCA). However, embedding algorithms of these works are designed for linear space, while intuitively network nodes might distribute on such nonlinear manifold as a sphere. Therefore, it is critical to find a more proper embedding algorithm that can best preserve distances after they are mapped into low-dimensional Euclidean space. To address the challenge, this paper proposes a distance prediction method using ISOMAP to exploit the nonlinear global embedding. As Fig. 1 shows, ISOMAP is proposed to estimate distances between points by their geodesic distances (i.e., the length of shortest path) which is capable of exploiting the true low-dimensional space.

A. Embedding Network Distances

The framework of the proposed method is similar to GNP. First N landmark nodes $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N$ are selected. Using ISOMAP for network distance embedding, $\mathcal{L}_i$'s coordinate vector $\mathbf{x}^{\mathcal{L}_i}$ can be obtained by which the low-dimensional Euclidean space can be spanned. Meanwhile these coordinates can minimize the square error between $d_{\mathcal{L}_i\mathcal{L}_j}$ and $\hat{d}_{\mathcal{L}_i\mathcal{L}_j}$ [7]. The key procedure are as follows:

- Construct a weighted Graph G over the nodes, with weight $d_{\mathcal{L}_i\mathcal{L}_j}$ assigned to the corresponding edges.
- Use Floyd's algorithm to calculate the shortest path $d^G_{\mathcal{L}_i\mathcal{L}_j}$ between $\mathcal{L}_i$ and $\mathcal{L}_j$.
- Use MDS to find m-dimensional coordinates $\mathbf{X}_m = [\mathbf{x}_m^{\mathcal{L}_1}, \mathbf{x}_m^{\mathcal{L}_2}, \ldots, \mathbf{x}_m^{\mathcal{L}_N}]^T$ best preserving distance given the pairwise matrix $\mathbf{D} = (d^G_{\mathcal{L}_i\mathcal{L}_j})$:
  - calculate $\mathbf{B} = -\frac{1}{2}\mathbf{HDH}$ where $\mathbf{H} = \mathbf{I} - 1/N\mathbf{1}\mathbf{1}^T$ and $\mathbf{1} \in \mathbf{R}^m$.
  - decompose $\mathbf{B} = \mathbf{V}\mathbf{\Gamma}\mathbf{V}^T$, where $[\mathbf{V}, \mathbf{\Gamma}] = eig(\mathbf{B})$.
  - $\mathbf{X}_m = \mathbf{V}_m\mathbf{\Gamma}_m^{1/2}$, where $\mathbf{V}_m$ and $\mathbf{\Gamma}_m$ are the top m eigenvectors and eigenvalues respectively, and $\mathbf{x}_m^{\mathcal{L}_i}$ is the i_{th} row of $\mathbf{X}_m$.

When it comes to the dimension of the embedded Euclidean space, most of previous methods need determine it experimentally which limits their adaptability and practicality, while in this paper MDS is applied to exploiting the dimension size of embedding Euclidean space according to distance preservation. For $m = 1, 2, \ldots, N$, $\mathbf{X}_m$ can be calculated in MDS procedure, and then $\hat{d}_{\mathcal{L}_i\mathcal{L}_j} = \|\mathbf{x}_m^{\mathcal{L}_i} - \mathbf{x}_m^{\mathcal{L}_i}\|_2$, where $\|\cdot\|_2$ is L_2 norm in Euclidean space. Therefore we can choose the best and lowest dimension m^* by minimizing the square error:

$$m^* = \min\{\arg\min_{m} {}_{\mathcal{L}_i, \mathcal{L}_j \in \{\mathcal{L}_1, .., \mathcal{L}_N\} | i>j} (\hat{d}_{\mathcal{L}_i\mathcal{L}_j} - d^G_{\mathcal{L}_i\mathcal{L}_j})^2\} \qquad (3)$$

In the second part, for each non-landmark node or ordinary node $\mathcal{H}_i$, its coordinate can be obtained by optimizing error minimization as GNP does:

- Measure its distance $d_{H_i\mathcal{L}j}$ to landmarks $\mathcal{L}_j$.
- Calculate the shortest path distance $d^G_{\mathcal{H}_i,\mathcal{L}j}$ between $\mathcal{H}_i$ and each landmark $\mathcal{L}_j$ easily.
- Find the coordinate $\mathbf{x}_{m^*}^{\mathcal{H}_i}$ of $\mathcal{H}_i$ in m^*-dimensional Euclidean space just by minimizing the relative error between measured and predicted distances as GNP:

$$\mathbf{x}_{m^*}^{\mathcal{H}_i} = \arg\min \sum_{\mathcal{L}_j \in \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N\}} \left(\frac{\hat{d}_{\mathcal{H}_i\mathcal{L}_j} - d^G_{\mathcal{H}_i\mathcal{L}_j}}{d^G_{\mathcal{H}_i\mathcal{L}_j}}\right)^2 \qquad (4)$$

Where $d_{\mathcal{H}_i,\mathcal{L}_j}$ and $\hat{d}_{\mathcal{H}_i,\mathcal{L}_j}$ are the measured and predicted distances between $\mathcal{H}_i$ and $\mathcal{L}_j$ respectively.

B. Predicting Distances

After the coordinate of each node are known, the distance between any two nodes can be predicted by

$$\hat{d}_{\mathcal{H}_i,\mathcal{L}_j} = \|\mathbf{x}_{\mathcal{H}_i} - \mathbf{x}_{\mathcal{L}_j}\|_2 \qquad (5)$$

Our method shares a similar framework proposed by GNP. It has been discussed in GNP that this framework has high scalability and the peer-to-peer architecture is easy to deploy. As to complexity, for ordinary nodes the evaluation of objective function (4) has identical time complexity. However, for landmark nodes GNP's iterative nonlinear optimization is computationally expensive and cannot guarantee the global optimal solution, while our global embedding method just conducts shortest path calculation and matrix decomposition whose time complexities are both $O(N^3)$ roughly. Furthermore, performance of iterative optimizing method depends on configuration of a number of parameters like step length, iteration number, and others, while our method using ISOMAP avoids the parametric complexity.

IV. Experimental Evaluation

In this section, we will introduce the data studied, describe implementation of GNP and the proposed method, and present and analyze experimental results.

A. Evaluation Methodology

1) Data: The data sets named Global Dataset and Abilene Dataset [2] contain 19 and 10 nodes acting as probes respectively. In each set, part of these probe nodes can be chosen as landmarks, and the rest together with 869 and 127 nodes respectively will be used as ordinary nodes.

2) Implementation: We have conducted a set of experiments using the data sets mentioned above to compare the proposed method with GNP. For GNP, the relative error is adopted as the objective function, because relevant research has already demonstrated that GNP performs well when such function is used [2].

In the experiments, BFGS (Broyden-Fletcher-Goldfarb-Shanno) [9] is used to optimize the objective functions. As we think, BFGS as a popular method used to solve unconstrained nonlinear optimization problems can find solutions of high quality and converges quickly. In order to objectively evaluate the methods, landmarks of both methods and initial points of BFGS are randomly selected. 100 experiments are conducted and all their relative errors $\frac{|\hat{d}_{\mathcal{L}_i,\mathcal{L}_j} - d_{\mathcal{L}_i,\mathcal{L}_j}|}{\min(\hat{d}_{\mathcal{L}_i,\mathcal{L}_j}, d_{\mathcal{L}_i,\mathcal{L}_j})}$ are averaged as the performance measurement in this paper.

B. Experimental Results

As discussed above, the proposed method applies MDS to determine the dimension of embedding Euclidean space according to (3). In fact, the optimal solution of (3) can be easily obtained since the error decreases monotonically as the dimension increases. This property can be illustrated by experiments on the error between predicted and measured distances, where dimension size varies and 15 landmarks are used on Global Dataset. Fig. 2 shows such trend of distance error with respect to the dimension.

Corresponding to Fig. 2, Table I presents details of the experiments. As the dimension increases, the error first decreases dramatically from 0.2810 to 0.0195, then gradually slows down, and finally level down 0.0188. According to the property of the error, the optimal solution m^* of (3) actually is the lowest dimension which reaches the smallest error. Here for this case $m^* = 8$.

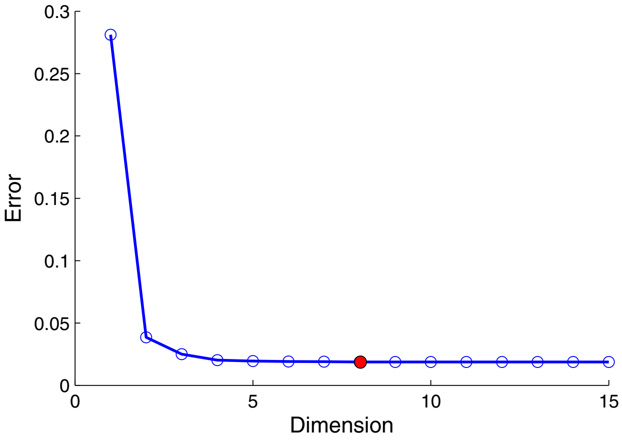

Fig. 2. Convergence with dimensions

TABLE I
Distance Square Error with Dimension

dimension	1	3	5	7	8	11	15
residual (10^{-2})	28.10	2.50	1.95	1.89	1.88	1.88	1.88

To validate the dimension determining mechanism, experiments on dimensions with respect to different landmarks are conducted.

TABLE II
Dimension Size with Landmarks

method	15 Landmarks	12 Landmarks	9 Landmarks	6 Landmarks
GNP	7	7	5	5
Proposed	10	8	7	4

For the proposed method, landmarks are randomly selected and the value for each number of landmarks in Table II is the average of all m^* in 1000 experiments. Table II indicates that dimension choice of the proposed method is almost consistent with that of the best performance in [2]. Therefore, it is reasonable to conclude that the dimension determining mechanism can help to find the best low-dimensional embedding Euclidean space.

Next, we turn our attention to distance prediction accuracy of the proposed method. We have conducted a set of experiments to compare the proposed method with GNP, where for the proposed method the dimension number is determined according to the distance error information at MDS step as discussed above, while for GNP it is set to 7 for 15 landmarks on Global Dataset and 8 for 9 landmarks on Abiline Dataset which achieved the best performance [2]. Fig. 3 and Fig. 4 show the average accuracy performance of both methods on

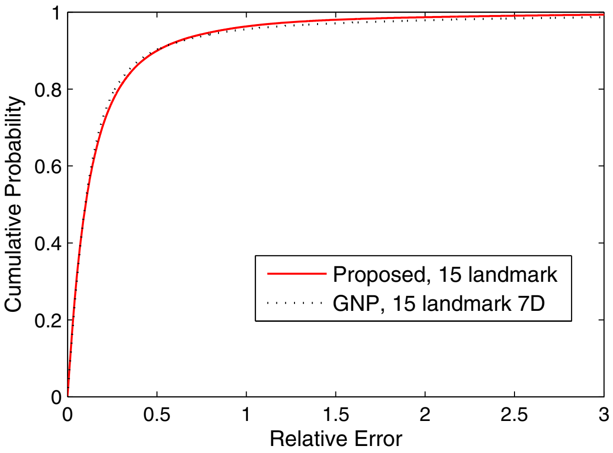

Fig. 3. Relative error comparison (Global)

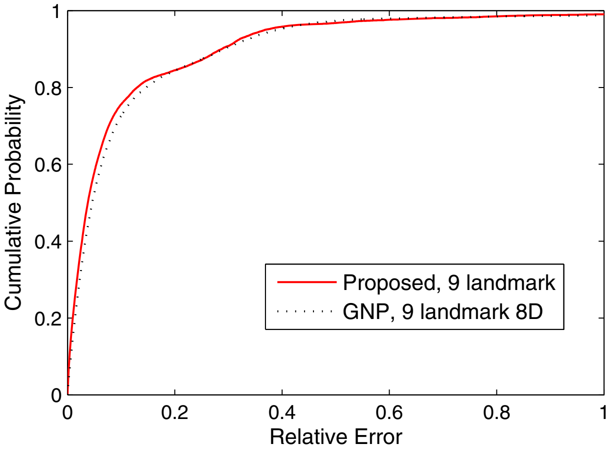

Fig. 4. Relative error comparison (Abilene)

Global and Abilene Dataset respectively.

From the figures, we can see that the proposed method achieves at least the same accuracy as GNP does on both data sets with respect to the relative error, while actually it performs better on Abilene Dataset: the cumulative probability of the proposed method reaches the 90% more quickly than GNP with lower relative error less than 0.5. Hence it can be inferred that the proposed method behaves accurately with respect to the average accuracy.

V. Conclusion

In this paper, we analyze network distance prediction methods in the literature and point out three problems unsolved in these methods: mapping nonlinear space into Euclidean space, determining dimension size of embedding Euclidean space, and reducing parametric and time complexity.

To solve these problems, a new distance prediction method using ISOMAP has been proposed. Since the geodesic distance is capable of revealing the true low-dimensional geometry of the manifold, substituting the shortest path distance for the distance between nodes helps to reduce the dimension and to find the best embedding low-dimensional space. Also, the proposed method employs MDS to calculate landmarks' coordinates through matrix decomposition. Applying MDS not only avoids iterative optimizing process and then reduce the parametric and time complexity, but also provides information of error between predicted and measured distances, according to which the dimension size can be determined easily. Discussion and experiments show that the proposed method is simpler, faster and performs more accurately than GNP.

Acknowledgment

The authors would like to thank anonymous reviewers for their valuable suggestions. They also want to thank Ph.D candidate Wei Su of Beihang University for discussions about optimization methods.

References

[1] P. Francis, S. Jamin, C. Jin, D. Raz, Y. Shavitt, and L. Zhang. Idmaps: A global internet host distance estimation service. IEEE/ACM Trans. on Networking, 9(5):525C540, 2001. 282, 288

[2] T. S. Eugene Ng and Hui Zhang. Predicting internet network distance with coordinates-based approaches. In IEEE INFOCOM, 2002

[3] F. Dabek, R. Cox, F. Kaashoek, and R. Morris. Vivaldi: a decentralized network coordinate system. In SIGCOMM, August 2004.

[4] M. Pias, J. Crowcroft, S. Wilbur, T. Harris, and S. Bhatti. Lighthouses for scalable distributed location. In IPTPS, 2003.

[5] L. Tang and M. Crovella. Virtual landmarks for the Internet. In Internet Measurement Conference, Miami, Florida, October 2003.

[6] J. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, vol. 290, pp. 2319C2323, 2000.

[7] T. Cox and M. Cox. Multidimensional scaling. in London: Chapman and Hall, 1994.

[8] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. Technical report, Department of Computer Science and Engineering, Michigan State University, 2006.

[9] Nocedal, Jorge; Wright, Stephen J. Numerical Optimization (2nd ed.), Berlin, New York: Springer-Verlag, ISBN 978-0-387-30303-1, 2006.