Community Graph Sequence with Sequence Data of Network Structured Data

Takehiro Yamaguchi*, Ayahiko Niimi†
*Graduate School of Systems Information Sciences, Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041-8655, Japan
email: g2109046@fun.ac.jp
†School of Systems Information Sciences, Future University-Hakodate
email: niimi@fun.ac.jp

***Abstract*—Recently, there has been increasing interest in data analysis for network structured data. The network structured data is represented the relation between one data and other data by graph structure. There are many network structured data such as social networks, biological networks in the real world. In this study, we will analysis the network structured data that has dynamic relation and complex interact with each data. And, we will approach the problem that is to extract transition pattern from the history of temporal change in their network structured data. Especially, in this paper, we will apply community graph sequences to graph sequences of network structured data that has large-scale and complex changes, and propose the method of extracting transition pattern of network structured data. We used social bookmark data as the data streams of analysis object and verified that social bookmark data is the network structured data that has large-scale and complex change.**

I. INTORODUCTION

Recently, there has been increasing interest in data analysis for an increasing number of network structured data. The network structured data is represented the relation between one data and other data by graph structure. Each vertex of graph represents a data and there are existing the relation between two data that are connected by each edge[1]. Link structure of the Web, human relationship network in social networking service, and gene networks in biology were branded it as examples of network structured data. For example, each vertex represents user and each edge represents relationships with friends between two users in human relationship network in social networking service. The vertices not only represent users but also represent communities and other groups. In that case, the edges represent the relation between groups.

Many of these data change the network structure with time change. For example, in the case of human relationship network in social networking service, the network structure is changed by newly adding users or withdrawing. In the case of gene networks, the network structure is changed by newly acquiring gene, missing, or mutation. For this reason, these data are not static structured data. To analyze such network structured data that change with time is not only necessary to analyze graph structure at some point but also necessary to analyze dynamics of graph structure. We consider that find out dynamics of the data that have such network structure lead to network structure prediction with a future or predicting interaction between each vertex with a future in network structured data.

In this paper, we will approach the problem that is to extract transition pattern from the history when it was given of network structured data that have changed with time. We consider representing network structured data as graph sequences, and we represent these dynamics changing by changing number of vertex and edge in graph sequences. Therefore, we could treat the problem mentioned above as a problem of mining subsequence from graph sequences. There are some previous works that mine subsequence from graph sequence([2, 3]). However, network structured data that is analysis object in this study have following feature that is different from analysis object in previous works.

1) a large part of the structure changes between two successive graph in graph sequences
2) the denseness of every graph in graph sequences

For example of network structured data that have above feature, there are Web log and other data streams. The data streams treat how a large data records have changed with time are momentarily produced, accumulated, and used as data flow. Especially, the data streams are a large data have the features that large amounts of data continue to arrive eternally over fast stream with temporally changing[4]. We consider that such data implements the feature of analysis object in this study by representing the relation in that as graph sequence. As far as we know, all previous work do not proposed the method of extracting transition pattern with graph sequences for these data.

In this paper, we will apply community graph sequences to graph sequences of network structured data that has large-scale and complex change, and propose the method of extracting transition pattern of network structured data with extending *GTRACE*[3]. We used social bookmark data as the data streams of analysis object and verified that social bookmark data is the network structured data that has large-scale and complex change.

II. RELATED WORKS

It was proposed *Dynamic GREW*[2] and *GTRACE*[3] as the method that mine subsequences from graph sequences. *Dynamic GREW* is the method that mine subsequence from

graph sequence that do not change number of vertex by representing insertion and deletion of edges as bit string. *GTRACE* is the method that enumerate subsequence from graph sequence that not only change number of edge but also change number of vertex by representing difference between two successive graph in a graph sequence with transformation rules. Moreover, it introduces a union graph that defines relevant vertices in graph sequences, and mine subsequence that include relevant vertices and exclude disrelated vertices.

The network structured data that is analysis object in this study could not apply *Dynamic GREW* because these data not only change number of edge but also change number of vertex in graph sequences. In graph sequence that is analysis object of *GTRACE*, there are following some requirement.

1) the number of vertices and edges in a graph sequence increase or decrease
2) change labels of vertices and edges in a graph sequence
3) a small part of the structure only changes between two successive graph in a graph sequence
4) each graph in a graph sequence is sparse graph

However, graph sequences of the network structured data that is analysis object in this study did not apply *GTRACE* because these data have large-scale and complex changes in sequences. In this paper, we propose the method that is extracting transition pattern of network structured data that have large-scale and complex changes in sequences.

III. PROPOSED METHOD

In this section, we will describe the method of extracting transition pattern from within the history of network structured data that complicated and changed large part of the structure in graph sequences. Our proposed method is extending *GTRACE (Graph TRAnsformation squenCE mining)*[3] to apply that network structured data. And, we will define the difference between our proposed method and *GTRACE*.

A. GTRACE

At first, we introduce forms to represent graph sequence in *GTRACE*. A labeled graph g is represented as $g = (V, E, L, f)$, where $V = \{v_1, v_2 \cdots, v_z\}$ is a set of vertices, $E = \{(v, v')|(v, v') \in V \times V\}$ is a set of edges, and L is a set of labels such that $f : V \cup E \rightarrow L$. $V(g)$, $E(g)$, and $L(g)$ are sets of vertices, edges and labels of g, respectively. An observed graph sequence is represented as $d = \langle g^{(1)}, g^{(2)}, \cdots, g^{(n)} \rangle$, where the superscript integer of each g represents the ordered step of the observation. For example, A graph $g^{(j)}$ is the j-th labeled graph in the sequence, $g^{(1)}$ is head of graph sequence, and $g^{(n)}$ is tail of graph sequence. Each vertex v of graph have a unique ID(Index), and it represented as $id(v)$.

To compactly represent a graph sequence, it represent interpolating two successive graph in graph sequence $g^{(j)}$ and $g^{(j+1)}$ as $s^{(j)} = \langle g^{(j,1)}, g^{(j,2)}, \cdots, g^{(j,m_j)} \rangle$, where $g^{(j,1)} = g^{(j)}$ and $g^{(j,m_j)} = g^{(j+1)}$. The observed graph sequence d represented by the interpolations as $d = \langle s^{(1)}, s^{(2)}, \cdots, s^{(n-1)} \rangle$. The interpolations is the order of graphs in the artificial interpolation and there can be various interpolations between the graphs $g^{(j)}$ and $g^{(j+1)}$.

The *GTRACE* extracts the interpolations by taking ones having the shortest length in terms of graph edit distance to reduce both computational cost and spatial cost of graph sequence mining. It defines a transformation of a graph by one of insertion, deletion and relabeling of a vertex or an edge be a unit, and the interpolations represented by the six TRs(Transformation Rules). That is, the graph sequence $s^{(j)} = \langle g^{(j,1)}, g^{(j,2)}, \cdots, g^{(j,m_j)} \rangle$ to interpolate $g^{(j)}$ and $g^{(j+1)}$ represented by TRs as $seq(s^{(j)}) = \langle tr^{(j,1)}_{[o,l]}, tr^{(j,2)}_{[o,l]}, \cdots, tr^{(j,m_j-1)}_{[o,l]} \rangle$, where tr is a transformation type which is either insertion, deletion, or relabeling of a vertex or an edge, o is the unique ID of a vertex or edge which the transformation is applied to, and l is a label to be assigned to the vertex or the edge by the transformation. The following is the six TRs.

1) Vertex Insertion : $vi^{(j,k)}_{[u,l]}$
   Insert a vertex having a label l and a unique ID u into $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
2) Vertex Deletion : $vd^{(j,k)}_{[u,\bullet]}$
   Delete an isolated vertex having a unique ID u in $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
3) Vertex Relabeling : $vr^{(j,k)}_{[u,l]}$
   Relabel a label of vertex having a unique ID u in $g^{(j,k)}$ to be l to transform to $g^{(j,k+1)}$.
4) Edge Insertion : $ei^{(j,k)}_{[(u_1,u_2),l]}$
   Insert an edge having a label l between 2 vertices having unique IDs u_1 and u_2 into $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
5) Edge Deletion : $ed^{(j,k)}_{[(u_1,u_2),\bullet]}$
   Delete an edge between 2 vertices having unique IDs u_1 and u_2 in$g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
6) Edge Relabeling : $er^{(u_1,u_2)}_{[u,l]}$
   Relabel a label of an edge between 2 vertices having unique IDs u_1 and u_2 in $g^{(j,k)}$ to be l to transform to $g^{(j,k+1)}$.

Thus, the interpolations $d = \langle s^{(1)}, s^{(2)}, \cdots, s^{(n-1)} \rangle$ represented by transformation sequence $seq(d) = \langle seq(s^{(1)}), seq(s^{(2)}), \cdots, seq(s^{(n-1)}) \rangle$.

GTRACE is the method to mine transformation subsequence $seq(d')$ that is subsequence of transformation sequence $seq(d)$ from a set of graph sequence $DB = \{d_i|d_i = \langle g_i^{(1)} \cdots g_i^{(n_i)} \}$. Especially, a support value $\sigma(seq(d'))$ of a transformation subsequence $seq(d')$ is defined as

$$\sigma(seq(d')) = \frac{|\{d_i|d_i \in DB, seq(d') \sqsubseteq seq(d_i)\}|}{|DB|},$$

and it to enumerate all transformation subsequence whose support value is greater than or equal to a minimum support threshold σ'.

In addition, it defines a union graph of a graph sequence d as $g_u(d) = (V(g_u(d)), E(g_u(d)))$ such that

$$V(g_u(d)) = \bigcup_{j=1,\cdots,n} \{id(v)|v \in V(g^{(j)})\}$$

$$E(g_u(d)) = \bigcup_{j=1,\cdots,n} \{(id(v), id(v'))|(v,v') \in E(g^{(j)})\},$$

and it defines similarly to transformation sequence $seq(d)$. It extracts union graphs of transformation subsequence that are isomorphic with g_u as relevant subsequence. This is forms to represent graph sequence and outline of *GTRACE*.

B. Extended GTRACE

We will extract the relevant subsequence from the history of network structured data that has large-scale and complex change modified with *GTRACE*, and treat these extracted subsequence as the transition pattern of these history data. However, the history of network structured data have following features that is different from the intended graph sequence in *GTRACE*.

1) a large part of the structure changes between two successive graph in graph sequences
2) the denseness of every graph in graph sequences

Therefore, when we have applied *GTRACE* to these data, these computational cost will become outrageous. It is easy to mine very short transition pattern, but it is difficult to mine long transition pattern. Because we consider that long transition pattern include more interest rules than short transition pattern, we did not apply *GTRACE*. In this study, we will approach following steps to apply *GTRACE* to the network structured data that has large-scale and complex change.

1) mine community structure from each graph in graph sequence
2) create sequence of community graph, where each vertex represent community, and each edge the relation between two communities
3) identify transition of community from elements in the construction of each vertex, and apply extended transformation rules of *GTRACE* to community graph sequence
4) extract enumerated transformation subsequence as the transition pattern

We will apply *GTRACE* to community graph sequence, because we consider that community graph sequence that have created from the graph sequence based on the history of network structured data that has large-scale and complex changes will become resembling the graph sequence that is analysis object in *GTRACE*. However, transformation pattern in community graph sequence exceed the range of transformation rules that is defined by *GTRACE*. We will extract these macro transition pattern from the history of network structured data that has large-scale and complex change by applying extended transformation rules of *GTRACE* to community graph sequence.

IV. EXTRACT TRANSITION PATTERN FROM COMMUNITY GRAPH SEQUENCES

In this section, we will discuss the extension of *GTRACE*, and define extending transformation rules.

A. Applying Community Graph Sequences

Our method extract community structure from each graph $g^{(j)}$ in observed graph sequence $d = \langle g^{(1)}, g^{(2)}, \cdots, g^{(n)} \rangle$, and introduce community graph sequence $cd = \langle cg^{(1)}, cg^{(2)}, \cdots, cg^{(n)} \rangle$ by defining community graph $cg^{(j)}$ where each vertex represent community and each edge represent the distance between two communities. We defined the distance between two communities as the degree that there are edges between users consisting each community. The distance between two communities having many edges is near or the distance is far. Therefore, when there is a edge between user in one community and user in other community, the edge is inserted between each node in community graph sequence and it labeled depending on the distance.

The community in this study is defined by the subgraph having high link density. The subgraph is extracted by the algorithm that intends maximizing modularity in the divisional process of network. Modularity is the evaluation figure of divided network. It evaluates how the divisional result divided the entire network to a proper balance of cluster having link density. For example, when the entire network is divided to L communities with no overlap where is $V_1, V_2, \ldots, V_L$, modularity Q is defined by following formula.

$$Q = \sum_{l \in 1 \ldots L} Q_l = \sum_{l \in 1 \ldots L} (e_{ll} - a_l^2)$$

e_{ll} means existing probability of link within V_l and, when it thought separating link end as outgoing link and incoming link no matter what the graph is non-directed graph, a_l means existing probability on the entire network of a total of link end within V_l. Each probability is obtained by following formula.

$$e_{ll} = \frac{1}{2m} \sum_{i \in V_l} \sum_{j \in V_l} A(i,j)$$
$$a_l = \frac{1}{2m} \sum_{i \in V_l} \sum_{j \in V} A(i,j)$$

$A(i,j)$ return 1 when there is a link between node i and node j in adjacency matrix of network or it return 0. A total of link end within network is $2m$. Therefore, e_{ll} finds that each community have high link density and, when it regarded the entire network as a community and randomly divided, a_l has introduced as correction term getting down Q.

We will extract community with *CNM(Clauset-Newman-Moore)* method[8] that is one of the method miximizing modularity. The good feature of *CNM* method is that computational cost is far small and that of conditioning parameter is unnecessary. Therefore, the method could calculate the clustering result by given network structured data([5, 7]).

B. Extending of Transformation Rules

In community graph sequence, because of defining each vertex as community, we did not just apply six transformation rules of *GTRACE* to community graph sequence. In this study, we extend transformation rules of vertices as following five transformation rules based on pattern of alteration of community structure from moment to moment[7], and just apply

GTRACE on the transformation rules of edges. Therefore, we will represent transition of community graph sequence with following eight transformation rules.

1) Vertex Generation : $vg^{(j,k)}_{[u,l]}$
   Insert a vertex having a label l and a unique ID u into $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
2) Vertex Disintegration : $vd^{(j,k)}_{[u,\bullet]}$
   Delete an isolated vertex having a unique ID u in $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
3) Vertex Relabeling : $vr^{(j,k)}_{[u,l]}$
   Relabel a label of vertex having a unique ID u in $g^{(j,k)}$ to be l to transform to $g^{(j,k+1)}$.
4) Vertex Integration: $vi^{(j,k)}_{[(u_1,u_2)\rightarrow u_3,l]}$
   Delete isolated vertices having a unique ID u_1 and u_2 in $g^{(j,k)}$ and insert a vertex having a label l and a unique ID u_3 into $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
5) Vertex Separation: $vs^{(j,k)}_{[u_1\rightarrow(u_2,u_3),(l_1,l_2)]}$
   Delete an isolated vertex having a unique ID u_1 in $g^{(j,k)}$ and insert a vertex having a label l_1 and a unique ID u_2 and a vertex having a label l_2 and a unique ID u_3 into $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
6) Edge Insertion : $ei^{(j,k)}_{[(u_1,u_2),l]}$
   Insert an edge having a label l between 2 vertices having unique IDs u_1 and u_2 into $g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
7) Edge Deletion : $ed^{(j,k)}_{[(u_1,u_2),\bullet]}$
   Delete an edge between 2 vertices having unique IDs u_1 and u_2 in$g^{(j,k)}$ to transform to $g^{(j,k+1)}$.
8) Edge Relabeling : $er^{(u_1,u_2)}_{[u,l]}$
   Relabel a label of an edge between 2 vertices having unique IDs u_1 and u_2 in $g^{(j,k)}$ to be l to transform to $g^{(j,k+1)}$.

We extract frequent transition pattern by applying extending above transformation rules of *GTRACE* to the network structured data that is analysis object in this study. But there is the problem how we identify each community between $cg^{(j)}$ and $cg^{(j+1)}$ in community graph sequence. We will define ID of each node in community graph sequence as set of elements constructing the community or set of unique ID in observed graph sequence. We will analysis set of elements constructing each community between $cg^{(j)}$ and $cg^{(j+1)}$ and identify each community depending on threshold.

V. EXPERIMENT

In this section, we will define interest relationship network in social bookmark data as network structured data of analysis object in this study, and describe the experiment that verified that these data have large-scale and complex changes.

A. Experiment Setting

The network structured data that is analysis object in this study is social bookmark data. Social bookmark is web service releasing own bookmarks on web and sharing with general public. Most typical and well-known examples are delicious[9], Hatena Bookmark[10], and livedoor clip[11]. We consider that it could apply analysis of network structured data of the relationship between each users in social bookmark to information recommendation bacause, in resent years, social bookmark went over basic concept sharing with users of social bookmark service and was changed "nearly interactive service" that advance communication between users of social bookmark[12].

We use EDGE Datasets[13] that is research datasets of livedoor clip. The datasets include about $1,500,000$ of bookmark data in about $25,000$ of livedoor clip users. In this study, we define interest relationship network in social bookmark with the datasets. The network represent each vertex of graph as user and represent each edge as the relation bookmarking same web page. We will represent time changes of this network structured data as graph sequence. This graph sequence is sequence of network structured data in users having different usage frequency and we forecast that these relation structure change with change of each user interest. And, social bookmark could find number of user bookmarking each web page. This could assume that "number of bookmark in each web page equal the popularity stakes in each web page" and not only use bookmarking web page but also use searching topic web page at the moment[14]. Therefore, we consider that above defined network have linked between many users and each graph in graph sequence is dense graph. In this paper, we will verify the number of vertices and edges and these amounts of change in graph sequence of our defined network.

B. Experimental Result and Consideration

In EDGE Datasets, we created two types of sequence from $1,568,833$ bookmark data in $25,356$ users from January 2005 to September 2009. One is weekly graph sequence composed of the graph created by divided social bookmark data by week, and the other is monthly graph sequence composed of the graph created by divided social bookmark data by month. We focused on the difference between two successive graph in each sequence and defined number of vertex insertion, vertex deletion, edge insertion, and edge deletion as vi, vd, ei, and ed respectively. Because we will verify that there are the differences of features of dynamics in two sequences, we created two type of sequence. We calculate there number as number of dynamics, and show number of vertex dynamics of weekly sequence, edge dynamics of weekly sequence, vertex dynamics of monthly sequence, and edge dynamics of monthly sequence in Fig.1, Fig.2, Fig.3, and Fig.4 respectively.

In Fig.1, Fig.2, it not only showed large amount of changed vertices and edges from about 79th graph in weekly sequence and but also showed that amount of inserted vertices and edges are in the same range amount of deleted vertices and edges. In Fig.3, Fig.4, it showed getting the same results from about 18th graph in monthly sequence. Therefore, it showed that a large part of the structure changes between two successive

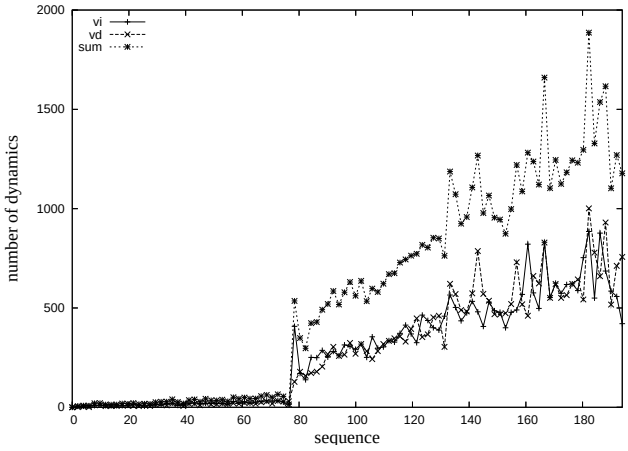

Fig. 1. Dynamics of Vertex in Weekly Sequences

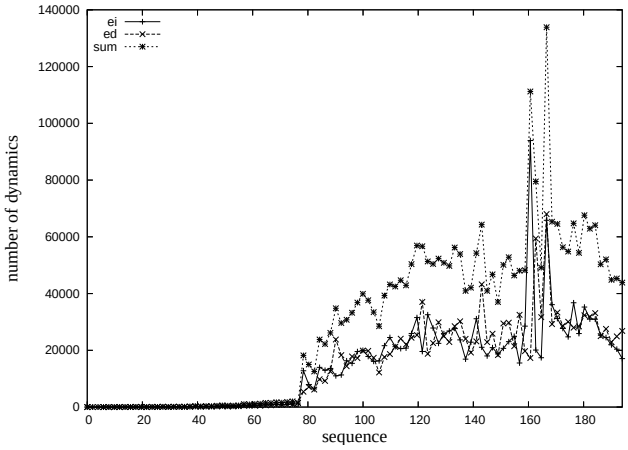

Fig. 2. Dynamics of Edge in Weekly Sequences

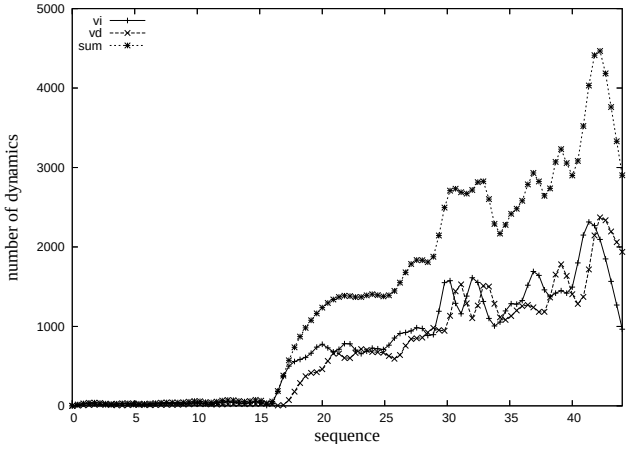

Fig. 3. Dynamics of Vertex in Monthly Sequences

graphs in graph sequence. Because 79th graph in weekly sequence and 18th graph in monthly sequence included June 27, 2006 when livedoor clip was released, we consider that the

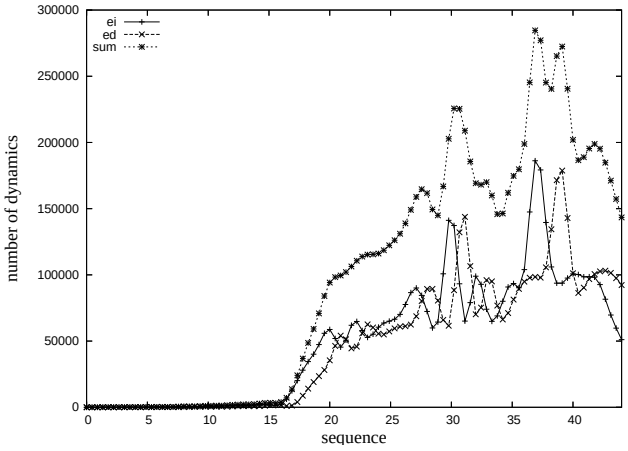

Fig. 4. Dynamics of Edge in Monthly Sequences

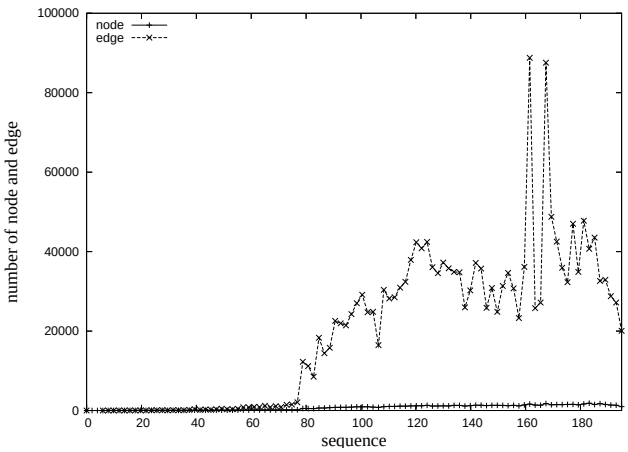

Fig. 5. Number of Vertex and Edge in Weekly Sequence

interest relationship network became the network having large-scale and complex changes. Comparing dynamics in weekly sequence and dynamics in monthly sequence, it showed that monthly sequence have big dynamics than weekly sequence and steep change of the number of dynamics in weekly sequence, compared with gradual change of the number of dynamics in monthly sequence. And, we show number of vertices and edges in weekly sequence and number of vertices and edges in monthly sequence in Fig.5 and Fig.6 respectively.

In Fig.5 and Fig.6, it showed that number of edges is larger than number of vertices from about 79th graph in weekly sequence and about 18th graph in monthly sequence respectively. Therefore, it showed that each graph in graph sequence is dense graph.

In these circumstances, graph sequence of interest relationship network in social bookmark is the data that has large-scale and complex changes. And, in the future task, because two created sequences differed in amount of dynamics and the process of dynamics, we will verify what extracted community and transition pattern of community from these sequences.

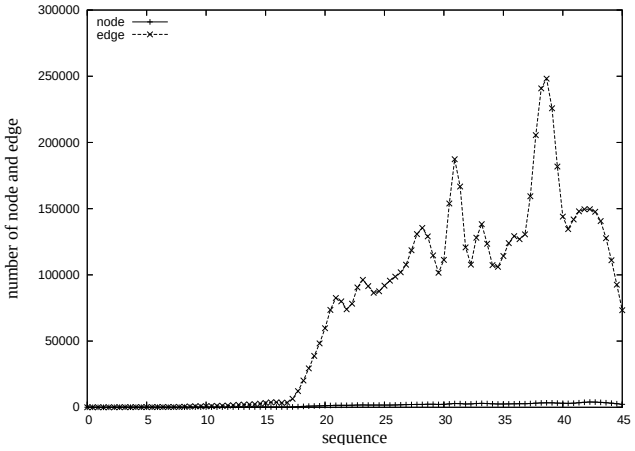

Fig. 6. Number of Vertex and Edge in Monthly Sequence

VI. CONCLUSION

In this paper, in the problem that is extracting transition pattern from given the network structured data that has large-scale and complex changes, we proposed introducing community graph sequence to graph sequence that is analysis object in this study and extracting method of transition pattern with extending transformation rules of *GTRACE* . We used social bookmark data as the data streams of analysis object and defined interest relationship network in social bookmark data. We verified weekly sequence and monthly sequence, and obtained the result that graph sequence of interest relationship network in social bookmark data is the data having large-scale and complex changes. In future research, we will apply transition pattern extracting from graph sequence to predicting network structured data.

REFERENCES

[1] H. Kashima: Survey of Network Structure Prediction Methods, *Journal of Japanese Society for Artificial Intelligence*, Vol.22, No.3, pp.344-351, 2007.
[2] K. Borgwardt, H. Kriegel, and P. Wackersreuther: Pattern Mining in Frequent Dynamic Subgraphs, *Proc. 2006 IEEE Int. Conf. on Data Mining*, pp.818-822, 2006.
[3] A. Inokuchi, and T. Washio: A Fast Method to Mine Frequent Subsequence from Graph Sequence Data, *Proc. 2008 IEEE Int. Conf. on Data Mining*, pp.303-312, 2008.
[4] H. Arimura: Recent Development of Mining Algorithms for Data Streams, *The transactions of the Institute of Electronics, Information and Communication Engineers*, D-I J88-D-I(3), pp.563-575, 2005.
[5] K. Yuta: Community Extraction Analysis and the Perspectives, *Operations research as a management science*, Vol.53, No.9, pp.529-535, 2008.
[6] M. Girvan and M.E.J. Newman: Community structure in social and biological networks, *PNAS*, Vol.99, No.12, pp.7821-7826, 2002.
[7] J. Ohwada, S. Yoshii, and M. Furukawa: Observing Change of Community Structure on Evolving Networks, *IPSJ Journal*, Vol.49, No.2, pp.765-773, 2008.
[8] A.Clauset, M.E.J. Newman, and C. Moore: Finding community structure in very large networks, *Physical Review* E, Vol.70, p.066111, 2004.
[9] delicious, http://delicious.com/
[10] Hatena Bookmark, http://b.hatena.ne.jp/
[11] livedoor clip, http://clip.livedoor.com/
[12] T. Nishitani: 0. Overview of SBM Study Group(⟨Special Feature⟩Towards a Revolution in SBM), *Journal of Information Processing Society of Japan*, Vol.49, No.12, pp.1410, 2008.
[13] EDGE Datasets, http://labs.edge.jp/datasets/
[14] M. Yokota: 1. Social Media and Marketing(⟨Special Feature⟩Towards a Revolution in SBM), *Journal of Information Processing Society of Japan*, Vol.49, No.12, pp.1411-1414, 2008.