Data Crystallizer

Tool for Discovering Unobservable Events

Yukio Ohsawa and Takaichi Itoh

1. Graduate School of Engigeering, The University of Tokyo (y.ohsawa@gmail.com) 2 Keio University

Abstract

It is only the observable part of the real world that can be presented in data. For such a scattered, i.e., an incomplete and ill-structured data, *data crystallizing* aims at presenting the hidden structure by inserting dummy items corresponding to unobservable, i.e., hidden events, to the given data on past events. The existence of hidden events and their position in the environment will be visualized as a result of data crystallizing. This basic method is expected to be applicable for various real world domains to which chance-discovery methods have been applied. This project aims at developing the process of data crystallizing, with a new tool extending KeyGraph, based on the process of chance discovery. In the research, experiments will be made using artificial data obtained from simulating the target of intelligence analysis, i.e., organized crimes. Then, the method will be applied to real workplaces with real data, real analysts, in real world domains.

1. Introduction

The basis of this proposal is Chance Discovery, which means to discover a *chance*, defined as an event significant for making a decision [1]. Using existing data in business and natural/social sciences, we have been achieving successful chance discoveries in various domains [2,3].

In the case of marketing, a customer's action to buy a product is an event to be dealt with in chance discovery. The co-occurrences among products in the order-cards of customers have been visualized as the map of the market, using a tool KeyGraph. In Nittobo Inc., the marketing team organized group meetings based on the Double Helix Process of chance discovery, with looking at the market map of KeyGraph [3,4], as illustrated late in this proposal. The company found the *bridges* between *islands*, where an island means a cluster of products popularly bought as one set, and a bridge means a product tending to be bought with products in multiple different islands. The company promoted the sales of products corresponding to the bridges, and it stimulated the customers to travel from/to islands across bridges. That is, customers found interesting new islands of products. This effect raised the sales performance of Nittobo Inc.

However, some events are darkly hidden, i.e., unobservable. As a result, the existing data may miss significant events for the decision of domain experts. This has been forcing them hurdles hard to overcome. For example, underground activities by criminal/terrorist groups may be really organized by management systems. Evidence Extraction and Link Discovery (EELD) progra [5,6,7], is leading the research trend of discovering significant links among items, on various data. The ultimate application domain of EELD can be *intelligence analysis*, where data of organized criminal behaviors are analyzed for capturing the signs of future crimes. A difficulty in intelligence analysis is that a group may be ruled by a hidden director, aiming at an ultimate goal such as a political revolution, or by a hierarchical system with hidden sub-managers in all levels. It is hard to notice such a fundamental bone of the organization, because the bone works silently in the most intelligent stage when the group members are planning new activities.

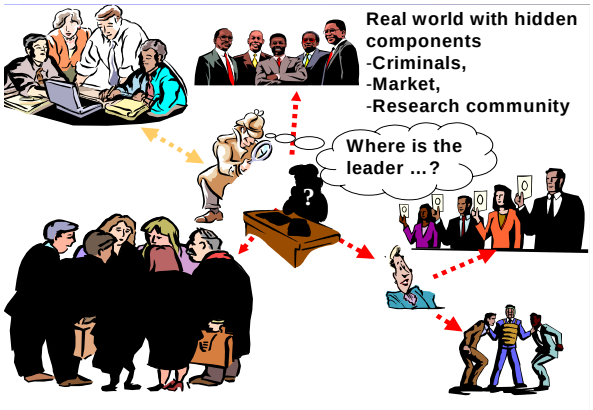


Fig.1 A hidden leader may connect the active group. He/she does not appear in the data.

In this project, we plan to develop a process called *chance discovery with data crystallization* for enabling to catch unobservable chances. Here, the user interacts with the data and additionally visualizes *dummy* events, not given in the data but should be assumed to have really occurred, together with the events in the existing data. This process is executed using a tool of *data crystallization* extended from KeyGraph (Section 2 and [4,8]). On its visual output, the user understands the underlying structures among both observable and unobservable events in the real world. This study is a basic research, meaning to extend the Double Helix (DH) process of chance discovery (in [1]), in order to cope with the stronger incompleteness of information in more complex real world than we dealt with in previous projects.

2. KeyGraph: Visualized scenario map for scenario communications

KeyGraph is a tool for visualizing a scenario map. If the environment here means the teamwork of a (e.g. criminal) group, KeyGraph shows the relation of members on their co-existing frequencies. This method has been applied to members of on-line [9] and real-world communities. For example, "`saru`" in Eq.(2) can be regarded as "`saru_attend`" i.e., an event that a member appeared in a meeting place. Here, let data D express a set of meetings, putting a period (``."``) at each end of meeting.

*D*1 =

```
tsubak
i       saru    ogura   kuwa.
                yoshid          kawa san
osawa   yuji    a       xu      i    o.
tsubak                  kawai
i       saru    kuwa     .
kawai   kuwa    nagai
        yoshid  tsubak
ogura   a       i       kawai xu.
….
```

(1)

KeyGraph, of the following steps, can then be applied to *D*1 ([4,8]). Fig.5 is the result.

***KeyGraph-Step* 1:** Events appearing many times in the data (e.g., "`member1`" in Eq.(2)) are depicted with black nodes, and each pair of such frequent events co-occurring often in the same set (i.e.., in the same set ending with a period) get linked with a black line. For example, `member1`, `member2`, and `member3` in Eq.(2) are connected with black lines in Fig.5. Each connected graph here forms one *island*, implying a basic context shared by the belonging members. If there are weak links, connecting islands which should be separated, they are deleted [8].

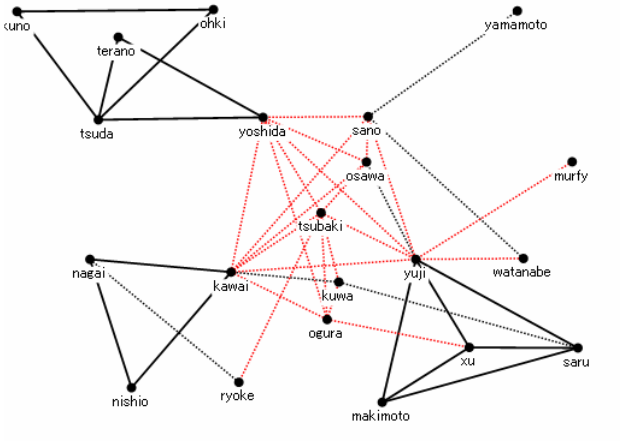


Fig.5. An example of *KeyGraph*: Islands are obtained from *D*1 in Eq.(2), including sets {yuij, xu, saru, makimoto} and {terano, tsuda, yoshida} etc. The nodes between islands show bridges, which may trigger the big teamwork of the whole group.

***KeyGraph-Step* 2:** Events co-occurring with multiple islands, e.g., `member9` in Eq.(2), are obtained as *hubs*. A path of links connecting islands via hubs is called a *bridge*. If a hub is rarer than the black nodes, it is colored in a different color (e.g. red or white) than black. We can regard such a new hub as a candidate of *chance*, i.e., events which may be rare but significant for context-jumping decisions.

3. Data Crystallizing as Extension of Chance Discovery

Data crystallization is different from the compensation for the missing data that has been earnestly studied in data mining. That is, the missing data in the previous context meant the loss of values for existing known variables. On the other hand, data crystallizing challenges a missed component corresponding to a variable rather than the value of a variable. That is, the user hardly knows any data is missed. If there is no glass (i.e., a variable), we can hardly notice wine (i.e., the value) should be served. Similarly, if there is no knowledge that the ninth item in a line of Table 1 may exist, we can hardly expect the item should be found, without a data crystallizing technique.

This project aims at developing the process and tools for chance discovery with data crystallizing. The sphere of real world applications linked from this

basic research is widened to include intelligence analysis, development of new products, aiding corporate behaviors by detecting unknown common interest of employees, etc.

4. The Method of Data Crystallization

The tool takes the following crystallizing steps.

[The procedure of data crystallization: Step C]
k :=1; *Hidden*_0 := {}; *line*_0 :={}; M_1 := a value given by the user;
for M_2 = 1 to M_1 (M_1 +1)/2 **do**
  for all $i, j \in \{0, 1, \ldots, N\}$ s.t j L.E. i **do**
    if line_i EQ. line_j **then** *Insert* (D, k, i, j);
  H := *KeyGraph*(D, M_1, M_2, M_3 := M_1/2);
  for j =1 to N **do**
    If $j \notin H$ **then** *Delete* (D, k, j);
  If $H \neq$ *Hidden_k*-1 **then**
    Hidden_k := H;
    for m =0 to k-1 **do**
      Delete (D, m, *Hidden_m* $\cap$ H);
      Hidden_m := *Hidden_m* \ H;
    k := k+1;

Here, D is the data to be analysed with KeyGraph in the function KeyGraph(D, M_1, M_2, M_3). N is the number of lines (co-occurrence units) in the data. *Insert* (D, k, i, j) means to insert k_j, the dummy node for the j-th line in the k-th level of crystallization, to the i-th line of data D and from data D. The second and the third lines of the procedure above mean to insert k_j to the j-th line, and, if there is a line (the i-th line) of the same set of items as the j-th line, k_j is inserted to all those lines. *Delete* (D, k, j) means to delete k_j, the dummy item for the j-th line in the k-th level, for all its appearances in data D. H represents the set of the line-numbers where the dummy items, which appeared on the bridges of the current KeyGraph, are positioned in the data. *Hidden_k* means the set of line-numbers with a dummy item crystallizing the data in the k-th level. line_j represents the set of items in the j-th line.

Informally, we can explain the process as follows. First, k, the value of crystallization level, is set to 1. The value of M_1 is defined by the user(s). Then, M_2 is incremented from 1, until as many black links as M_1 (M_1 +1)/2 appear, to connect all the nodes in one island. Then, dummy items are inserted to D. If two or more lines have the same set of items, the same dummy item is inserted to all those lines, suffixed with the line-number of the last of those lines.

To this data with inserted dummy nodes, KeyGraph is applied. Then, the newest (i.e., of level k) dummy items which did not appear on the bridges of KeyGraph get deleted from D. If a line includes more than one dummy item, all the dummy items in the line except for the highest level will be deleted. Here, the integer k, the level of crystallization, is incremented if some of the newest dummy nodes remain after this deletion. As a result, the following are obtained:

- A new data with dummy items, corresponding to hidden events connecting substructures in each level.
- KeyGraph(D, M_1, M_2, M_3) for the obtained data D, for an arbitrarily determined values of M_1, M_2, and M_3. By increasing the value of M_2, we can focus the output to the higher level of the hidden structure.

See Figure 9 for the result of crystallized KeyGraph for D1. The two figures show the results of data crystallization executed. The hierarchical structure underlying the data, including the dummy items, are visualized. By comparing the (a), (b), and (c) in Fig.9, it is clear that the structures in different levels of the same environment can be also seen by arranging the value of M_2. From (a), the user interprets hidden sub-leaders denoted by 1_x for integer x. From (b), 2_x shows hidden leaders in higher level. Increasing black links mode, (c) shows 3_2, hidden leader of the highest level.

Data crystallization works in the way like the crystallization of snow. A crystallizing item of the data plays the role like a particle of dust connecting molecules of water. The increase in M_2 corresponds to the decrease in temperature, so the gradual increase in M_2 makes a well-structured KeyGraph corresponding to a well-structured snow.

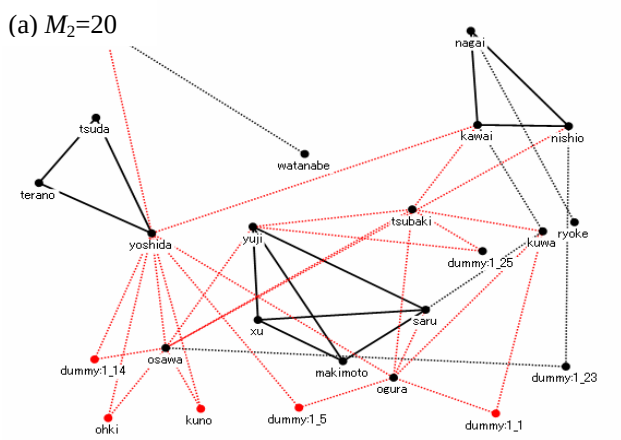
5. Concluding Remarks

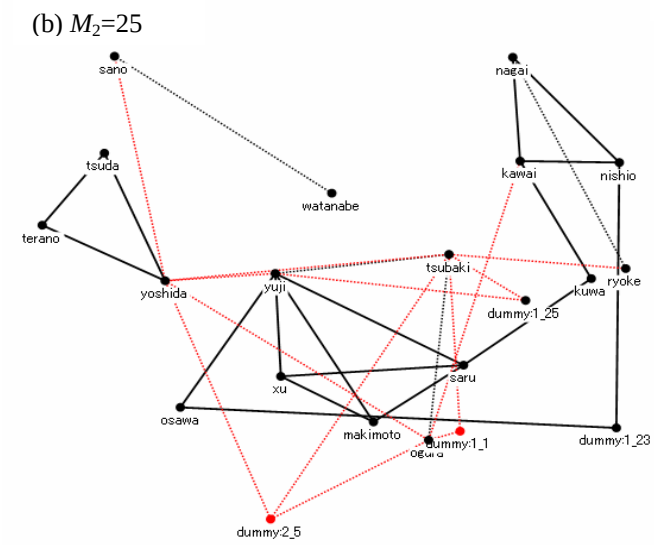
A tool of data crystallizing is presented. This means to extend chance discovery to the discovery of really studies on seen events in more uncertain environment than chance discovery have been dealing with. The sphere of real world applications linked from this basic research is widened to include intelligence analysis, where real leaders do not appear in such data as news about meetings.

Because the data crystallization is for understanding deep-level structures, dummy items cannot be understood if user is in an early stage of chance discovery. There is a risk of disturbing user's understanding if a complex structure is, so this tool should work only if user is concerned with deeper structures, in human's process of chance discovery.

For the time being, the studies in EELD is more oriented to automating discoveries or to coupling symbolic expressions of human knowledge with a machine learning system, whereas chance discovery

has been integrating the human process of externalizing the tacit experiences and the power of machines for finding a surprising trigger to the activation of the environment.





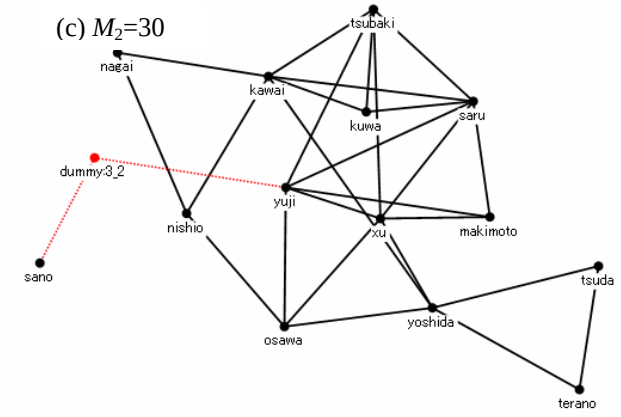


Fig.9 The crystallized KeyGraphs obtained with increasing the number of black linkes (M_2) by data crystallizer, for the data in Eq.(1). 2_5 is the dummy for higher level organizer than 1_1, and 3_2 is of even higher.

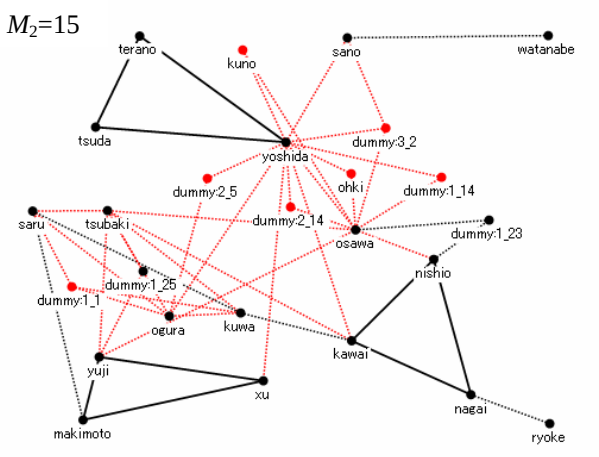


Fig.9 The crystallized KeyGraph,, i.e., KeyGraph for the data with final dummy items on M_2 of 15.

References

[1] Ohsawa, Y., 2003a, Modeling the Process of Chance Discovery, Ohsawa, Y. and McBurney eds, *Chance Discovery* pp.2—15, Springer Verlag (2003)

[2] Chance Discovery Consortium (CDC), http://www.chancediscovery.com (2004)

[3] Yukio Ohsawa and M. Usui: Workshop with Touchable KeyGraph Activating Textile Market, in *Readings in Chance Discovery*, Advanced Knowledge International (2005)

[4] Ohsawa Y, 2003b, KeyGraph: Visualized Structure Among Event Clusters, in Ohsawa Y and McBurney P. eds, 2003, *Chance Discovery*, 262-275, , Springer Verlag (2003)

[5] Senator, T., EELD Program, http://www.darpa.mil/ito/research/eeld/EELD_BAA.ppt (2001)

[6] Upal M.A., Performance Evaluation Metrics for Link Discovery Systems in Proceedings of the Third International Intelligent System Design & Applications, pages 273-282, Springer Verlag, (2003)

[7] Sutton, C., Brendan, B., Morrison, C., and Cohen, P.R., Guided Incremental Construction of Belief Networks. 5th International Symposium on Intelligent Data Analysis (2003)

[8] Ohsawa, Y., Benson, N.E., and Yachida, M., KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, Proc. IEEE Advanced Digital Library, pp. 12-18 (1998)

[9] Ohsawa, Y., Soma, H., Matsuo, Y., Usui, M., and Matsumura, N., Featuring Web Communities based on Word Co-occurrence Structure of Communications,

Proceedings of the Eleventh Conf. World Wide Web (WWW11), ACM (2002)