Modeling the Process of Chance Discovery by Chance Discovery on Double Helix

Yukio Ohsawa[1, 2] and Yumiko Nara[3]

1. Japan Science and Technology Corp, 2. The University of Tsukuba
3. Osaka Kyoiku University

Abstract

A case is presented for the double helical processing of chance discovery. Here, human and a data mining system *KeyGraph* co-work, each progressing spirally toward the creative reconstruction of ideas. The discovery of what we call *chances*, significant novel events, are to be realized in this process. The example shown here is its application to questionnaire analysis, for understanding the behaviors of the Internet users with discovering chances. Internet users are born and bred with face-to-face human relations in the real world, but their interactions with WWW are distilling new value-criteria, keeping personal real-world senses of rationality, ethics, etc. In our method for aiding chance discovery, base on the double-helix model, the in-depth interaction of the Internet, the personalities and the behaviors of people came to be understood with revealing unnoticed value-criteria. This lead to a new model of the process of chance discovery.

INTRODUCTION

$$Y=f(X) \qquad (1)$$

In Equation (1), Let function f define a deterministic system, which obtains a unique output Y for each input X. Even if a part of the process from the input to the output is non-deterministic, i.e., keeps multiple decision candidates, the overall process is deterministic if the final output is uniquely determined. For example, the process from the input of external information by a sensor of a robot, to its real action, is deterministic.

Here, let us put input X by an n-dimensional vector (x1, x2, ... xn) and consider to control Y indirectly by touching the values of X. With out loss of generality, we can denote the touchable part of X by $X0:(x_1, x_2, ...x_m)$ ($m<=n$). In case f is identified and $m=n$, we call f a programmed decision-making system. Otherwise, i.e., if f is not identified or $m<n$, we call f a non-programmed decision-making system (Simon, 1945). In previous system-identification methods, the targets has been programmed decision systems, especially in case f is not identified and $m=n$. This condition stands for artificial systems whose behaviors are determined by limited factors, due to the artificial design of input.

The process of human starting from information-acquisition to end in a determined decision or action is a deterministic system. However, this deterministic system is affected by various unknown factors, i.e., f is undefined and $m < n$ - the most hard to identify.

In previous social surveys, based on questionnaires asking a certain number of questions to a number of interviewees, they often applied statistic analysis (e.g. path analysis as AMOS) to the resultant data. This corresponds to introducing the approximation of $m=n$ to the target sample-set of people, for identifying f. Such an approximation might be valid if f is supported by a large sample set, because individual features of samples become ignorable or kill the influence of each other.

However, with the changes in the social infrastructures e.g. the growth of the Internet, the easy publication of individual ideas and dynamic flows of pia-to-pia information are increasing the impact of specific personalities onto the whole society. Socialists, market researchers, politicians, and educators need to be aware of specific answers implying emerging behaviors of subjects who are making decisions and actions on emerging opportunities and risks. In this paper we present *double-helix*, a model of the process of *chance* (an event, which might be rare, but is significant to decision making) *discovery*, applied to the survey of people's' behaviors on the Net and in the real world motivated by unknown factors. This lead us to modeling the process of chance discovery.

Double Helix: a Guideline of Chance Discovery

In the social survey, the unknown factors, i.e., (x_{m+1}, x_{m+2}, …x_n) are not registered as existing attributes of interviewees. No machine can tell what these factors really are, because the questionnaire does not include questions for unknown attributes. Especially if an answer is rare or novel, the organizer of a social survey normally does not know all factors motivating rare answers. For an educator and a market researcher, such a novel answer can be regarded as what we called a *chance* above, because newly identifying the unknown factors may help in promoting or reducing the unnoticed tendencies of students or customers.

So far, studies on Chance Discovery have been contributing to aiding human awareness of such chances, i.e., events significant for decision-making. For example, the relations between events in the target domain are obtained from data and visualized by *KeyGraph* (a method summarized later), and the group-wise exchange of ideas inspired by the graph enables the discovery (Fukuda and Ohsawa, 2001, Ohsawa, 2002). In this approach, few discussants' subjective awareness of the significance of rare events spread throughout the group. In (Ohsawa, 2002), the PUG criteria of a chance were proposed: *Proposabiliy* (ability to propose a certain decision/action based on the understanding of the chance), *Unnoticability* (difficulty to notice the significance of the chance), and *Growability* (the growth speed of the number of people who accept the proposal based on the chance). In the group-discussion approach, these criteria could be measured.

The success of group discussion for chance discovery inspired us to a revision from Fayyad's model of knowledge discovery process in (Fayyad, 1996). A key point for this revision is that it came to be considered that a chance can be something unnoticed (rare of new). In such a case, human cannot smoothly feel concerned with the chance because s/he is disturbed by cognitive dissonance (Festinger, 1957), due to the unfamiliarity of a rare event. Hence our model including the context-shifting cycles carrying human from *ignorance* to *concern*, *understanding* of the chance, and deterministic *decision/actions*, returning to a new ignorance of another chance, as depicted in Fig.1. The group of people stimulated steps of each other.

Awareness of unknown factors create hypotheses, and the observation validates and deepens the hypotheses. Because considering deepened factors helps in the new explanation of significance of chances, we become able to select significant events in the cycles (select 2). On the other hand, select 1 discards only noise which the human allows to discard, by the last cycle. Select 1 shrank in this manner, from Fayyad's, because events might have unnoticed significance.

In this paper, we show an integrated model of this phenomena and the process model in Figure 1 and show an exemplification of the process. This new model, called *double helix*, has two helical sub-processes as in Fig.2. One is the spiral process of chance discovery by human, which substantially the cycles in Figure 1 progressing to deeper awareness. The other helix is the process of computer(s), receiving and mining data ("DM" in Figure 1). The name "double-helix" means the parallel processing, i.e., the simultaneous runs of these pair of helixes (spiral processes) due to the input of "the subject(s)-data" monitoring the mind of the subject who aims to discover chances. For example, the words in the thought of discussants above form a subjects-data.

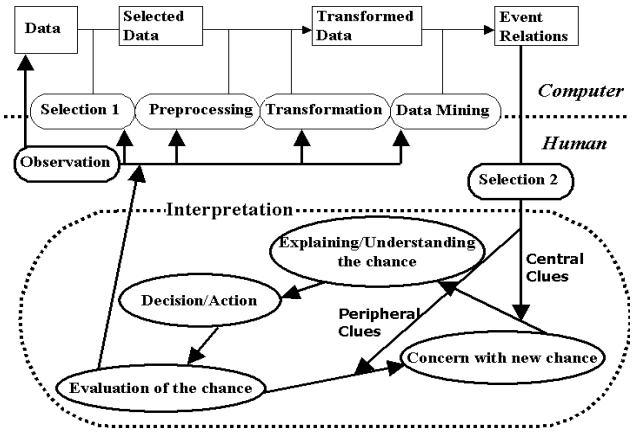

Figure 1. The model of Chance discovery process. In Fayyad's model, the interpretation and evaluation were put in one step, and observation did not appear.

This input occurs, while the subject is discussing or thinking – in previous data-mining, the computer was taking a rest while human were discussing about the last output for "the object(s)-data" that is the data for the target problem dealt with. Between these two helixes, interactions below occur:

- The subject-data is obtained from the subject's thinking process, bound for decision.
- The object-data is collected based on the subject's concern with the target domain
- The mining result from the subject-data (DM-a, in Figure 1) is reflected to the understanding the subject's understanding of the her/his own concern
- The mining result from the object-data (DM-b) is reflected to the understanding of the chance.

Application of Double Helix to understanding human behaviors on the web

In social sciences, the influence of the Internet to human behaviors is gathering attentions. With the process of chance discovery in Figure 2, we exemplify the human-computer interaction for understanding the influence of value-criteria of individual people to their behaviors.

The interaction includes data-acquisition with the questionnaire lead by Nara (Nara, 2001), supported by *KeyGraph* and the double-helix model having been developed by Ohsawa (Ohsawa, 2001) . The aim of this process is a chance discovery, where the chances are the novel and significant behaviors of Internet users and the reasons where the feelings come from. In this section, let us show a case the double helix lead us to reach an explanation of how chances in the dynamic human society make affection to the behaviors of Internet users.

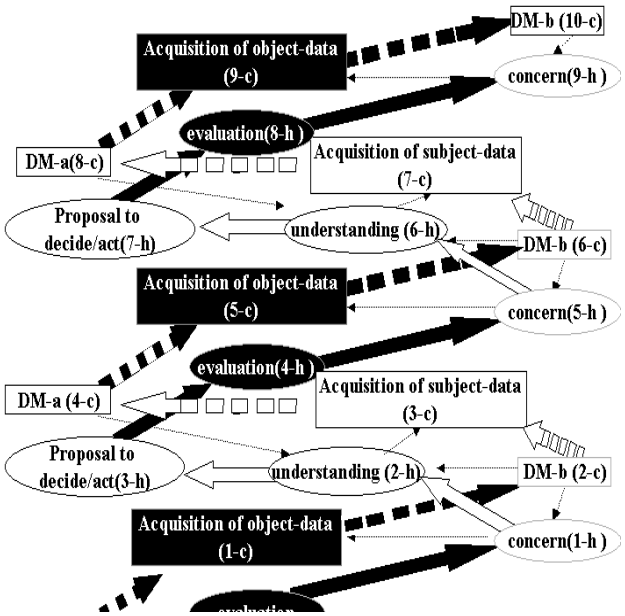

Figure 2 The double helical process of chance discovery (the numbers as "1-h" corresponds to each step below).

Hereafter, the number of each step with "-h" ("-c") appended, i.e., "*k*-h" ("*k*-c") for integer *k* means the information process of human (computer) in the increasing order of *k*. That is, the thick solid (dotted) arrows in Figure 2 depict the steps on the side of human(s) (computer). The thin arrows show the interaction between human and computer, itemized in the end of Section 2.

1-h) *Concerned* with behavioral factors of people

The first model by Nara was as in Figure 3, composed of five nodes and five links each corresponding to her hypotheses for explaining the behaviors of people in the real and the virtual (Net) world. In the figure, "general ethics" means the fundamental ethics of people affecting the behaviors both in the real (non-virtual) and the virtual world. On the other hand, the "information ethics" means the ethics reflected to the usage of information in the virtual world, e.g., the use of e-mails and production/protection of e-virus. "Empathy" is a personality relevant to ethics, defined as the ability to adopt the other person's position, to grasp his/her perceptions and feelings. In this figure, we find a sociologic way of thinking about behaviors of people, paying significant attentions to fundamental personalities.

1-c) The object-data: Questionnaire results

Based on Figure 3, Nara made questionnaires to university students in Japan. The students answered questions about their own behaviors and feelings about affairs in the real and the virtual world. Each question reflected each item in the nodes of Figure 3, the details being described in (Nara, 2001).

2-c) DM-b: Path analysis for the questionnaire data

In Figure 4, the hypothesized relationships between behaviors and personalities modeled in Figure 3 were mostly validated, by a path-analysis software (AMOS: www.smallwaters.com) applied to the questionnaire results. Here the usage of the Internet and empathy correlate with the ethics in question. (Nara, 2001) declared four of five hypotheses were validated.

2-h) *Understanding* of peoples' behaviors

The results in 2-c) form a summary of data and the validation of the relationships in Nara (author, and as the chance-discovery subject here)'s concern. By clarifying why these hypotheses came validated/rejected, one reaches an understanding of underlying causality. In this understanding, Nara mentioned "external sanctions" to people is clearer in the real world than in the virtual world. As a matter of fact, "sensation-seeking" people cannot be satisfied in the real world as easily as in the virtual world.

3-c) An article as a subject data of Nara

The subject-data we took here is paper (Nara, 2001), including the thoughts and the analysis above of Nara.

4-c) DM-a: the result of KeyGraph applied to the article

The paper above by Nara was processed by *KeyGraph* (Ohsawa and Yachida, 1999). Briefly, KeyGraph deals with data in the style as:

$$\begin{aligned} \text{Data} = {} & a1\ a2\ a3 \ldots a\mathit{m}. \\ & b1\ b2 \ldots b\mathit{n}. \\ & c1\ c2\ c3 \ldots c\mathit{p}. \\ & \ldots \end{aligned} \tag{2}$$

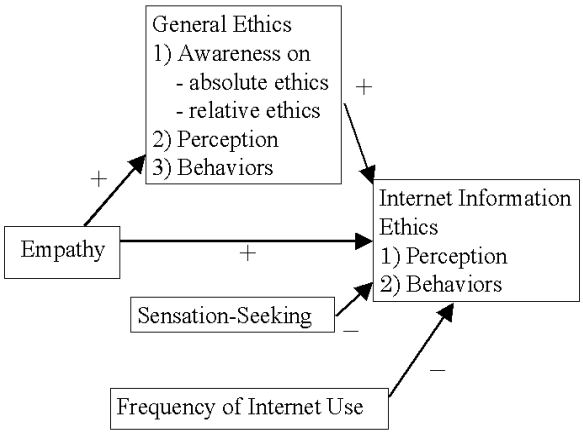

Figure 3. The initial model of the behaviors and personalities of people.

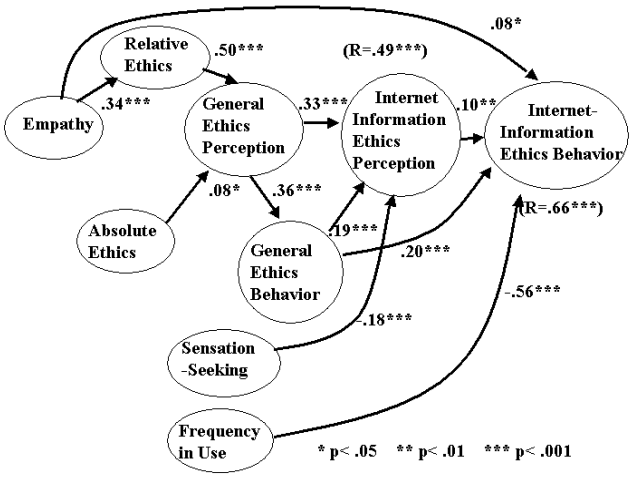

Figure 4. The validation of the model in Figure 3, employing the path-analysis software (AMOS).

Figure 5. The output of *KeyGraph* for paper (Nara, 2001).

In Eq.(2), each line ending with a delimiter (`.`) show the set of co-occurring items. If the data is a document, each line is a sentence composed of items corresponding to words in the sentence. If the data is a questionnaire-result data, each line is the answer-set of one examinee. *KeyGraph* obtains a set of clusters called foundations, each composed of frequent items depicted by nodes and the links between items of highest co-occurrences i.e., occurring in many same sentences. Each foundation represents a fundamental part of the data, in that they form the most familiar part of the data, for one who has some interest in the data domain.

Then *KeyGraph* shows rare (not so frequent as in foundations) items co-occurring with multiple foundations. These are candidates of "chances" we mean, i.e., rare but significant items carrying emerging significance in the environment or the self of the subject. In the case of a document, a chance is a concept carrying new assertions significant for the author's decision to write the document.

Here in 4-c), *KeyGraph* obtained Figure 5 from paper (Nara, 2001) processed as a document. In this figure, words for example as "sanction" and "scales" appeared as the candidates of chances. As we mentioned already, "sanction" is a key point newly emerging in the mind of Nara. "Scales" means the questions in the questionnaire for measuring the consciousness of each examinee. This implies the invariant concern of Nara to look for deepened factors of people and the "scales" for measuring the influence of those factors. This output made her aware of her own interests in these two items of which she, herself, was not explicitly conscious in writing the paper.

3-h) *Proposal* to realize ethical educations on the Net

Because an external sanction is harder to give in the virtual world than in the real world, Nara proposed ethical education enhancing internal sanctions in people's mind.

4-h) Evaluation of the results, and 5-h) new concerns

Although Nara became newly aware of her own concern with "sanction" and "scale," the model for/from the path-analysis did not include these factors. That is, her new interest to be reflected to the following analysis is summarized to look deeply at what makes people make extraordinary behaviors, if free from external sanctions.

5-c) New object-data: rare answers in the data.

The concern of Nara was, as mentioned above, leaning to the extraordinary manners in the real and virtual, e.g., crimes. In other words, she came to be interested in rare answers rather than frequent answers. Thus the object-data was revised to the answers by 50 or fewer of the total 493 examinees. In the framework of Fayyad, this is a reflection to the "selection" step. However, here we had an alternative choice to make a new questionnaire for young people in the prison if we have no obstacle for doing it – the acquisition of data is a generalized step of selection.

6-c) DM-b: *KeyGraph* applied to the set of few answers

To the rare-answer set in the results of questionnaire in 5-c), we obtained Figure 6 by *KeyGraph*. Here, individual examinee's set of answers was dealt with as a unit of co-occurrence i..e., between the delimiter mentioned in 4-c).

6-h) Understanding of Figure 6

Figure 7 depicts Nara's interpretation acquired after Figure 6. She noticed the significance of the concept "rational egoist tendency" by looking at Figure 6. This means the personality to make crimes if no penalty is prepared, matching with Nara's rising interest in sanctions. The effect of these visual data-mining is such externalization of subject 's unnoticed concerns.

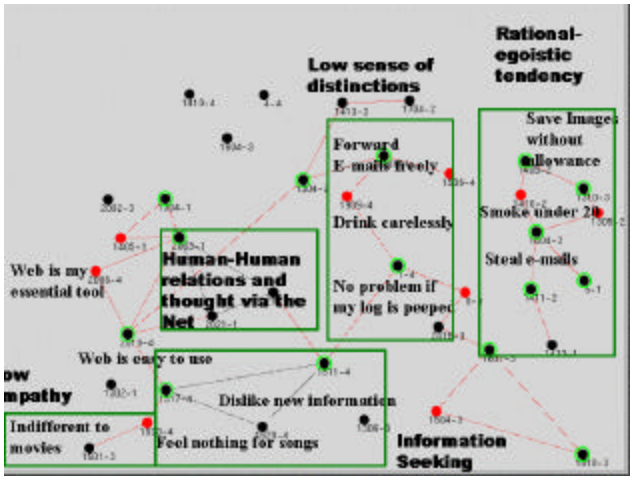

Figure 6. The result of *KeyGraph*, for the rare answers in the questionnaire.

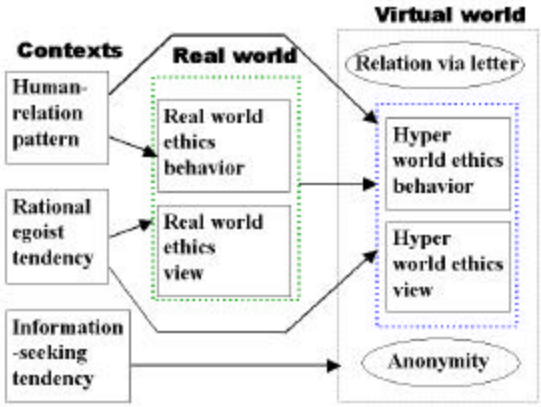

FIgure 7. A new model acquired from Figure 6.

7-c) The data of the subject's understanding of Figure 6, and 8-c) Its result by *KeyGraph*

Reference (Ohsawa. and Nara, 2001) described the study processed from 4-h) to 6-h). By *KeyGraph* we obtained Figure 8, showing rare but significant concepts as "deep-level ethical values" connected to "opinions of people's community." These concepts underlying in (Ohsawa. and Nara, 2001) were not noticed explicitly even by the authors, but fed back to the understanding, in 6-h).

7-h) New proposals

Considering the fundamental factors in Figure 6 to 8, i.e., (1) human relations in communities (2) information seeking and (3) deep-level ethics relevant to empathy and rational egoistic tendency, policies were proposed for the virtual society. For example, educators are desired to teach young people for acquiring empathy.

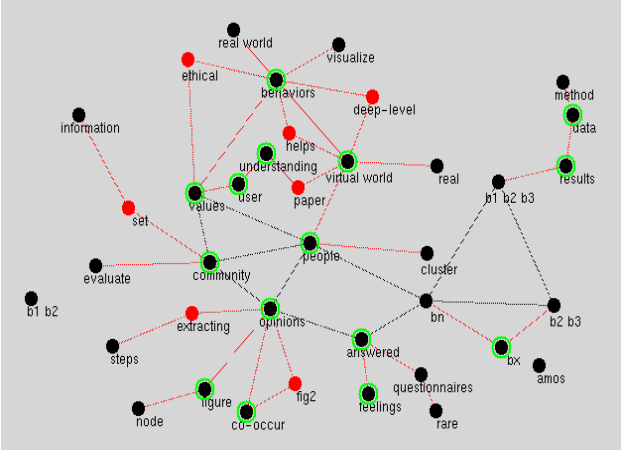

Figure 8. The output of *KeyGraph* for paper in (Ohsawa, and Nara, 2001)

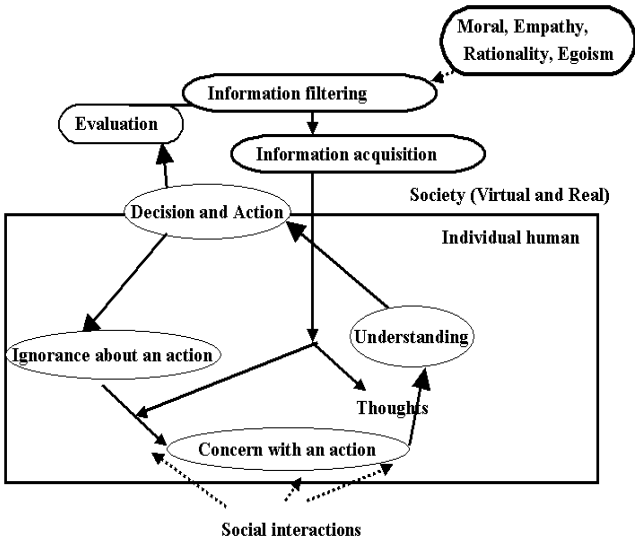

Figure 9. The decision-making model in the real and the virtual worlds, drawn on the board in the discussion.

8-h) evaluation and 9-h) new concerns

Although the proposals in 7-h) might be reasonable, it is still not realistic - seeking beneficial information can not be grasped from purely ethical points of view. Thus, we came to be concerned with business-sake information acquisitions and behaviors on the Internet and in the real world. Thus, we planned an inter-disciplinary discussion about decision making process on the Net, including us authors and other areas as business management, as mentioned below.

9-c) A merged data of subjects and objects: words in discussion, and 10-c) the output of *KeyGraph* as DM-a/b

We gathered a group of discussants to talk about the decision process with the information on the Net, with looking at Figure 6. This is a collection of object-data in that we had discussants who are also objects in the environment we are studying about. However, this is also a

collection of subject data in that we joined the discussion. Based on the lesson learned in (Fukuda. and Ohsawa, 2001) on group-working discovery of chances, Ohsawa called eight discussants from a variety of areas. That is, we had 3 experts of information technologies, 3 of management science, and us two. In the discussion, information and its relevance to real actions seemed to be the central topic. We reached Figure 9 as a model of decision-making.

Yet, the excitement during the three hours rather disturbed the discussants from systematic understanding of where their arguments were going to. We recorded the overall discussion and made a text file, as the data of subjects. Then, the text was processed by *KeyGraph*. Figure 10 was obtained as a result. As seen here (see the thick letters for English translation), the personality-factors made a significant bridge from the consciousness in human mind, to the embodied decisions and actions. That is, discussants paid strong attentions to empathy, rational egoism, and information seeking tendency, as significant factors affecting human decisions. The change from Figure 9 to 11 is a clear evidence of the effect of showing Figure 10.

10-h) Understanding and 11-h) Proposal

As a result, we obtained Figure 11 as the latest model of decisions on the Net, affected by personalities. The new proposals obtained are:

- Web-based (real-shop) marketing to customers after distinguishing their personalities whether they accept e-mail (face-to-face) recommendations or not.
- Multi-path (via the Net and face-to-face) negotiation in business with path-switching strategies.

Here we can also find that we may lose the point of a discussion during the time of exchanging opinions in complexly, and visualization of the overall conversation hours can aid the awareness of each discussant about the core ideas being talked about. Thus subject-data works.

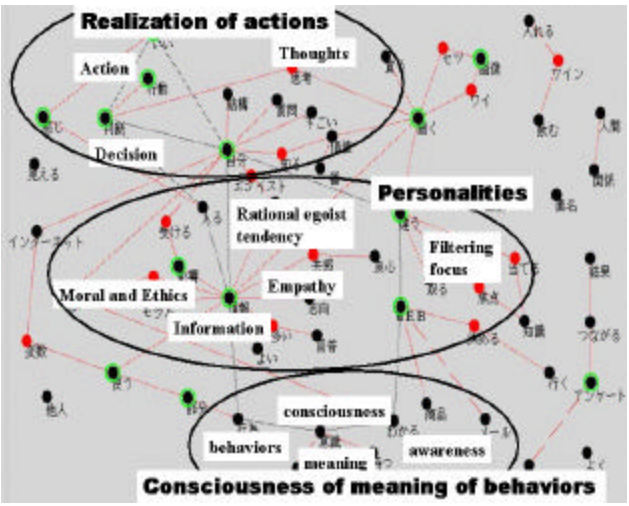

Figure 10. The result of *KeyGraph*, to the record of discussion with looking at Figure 6.

The Final Revision of the Process Model

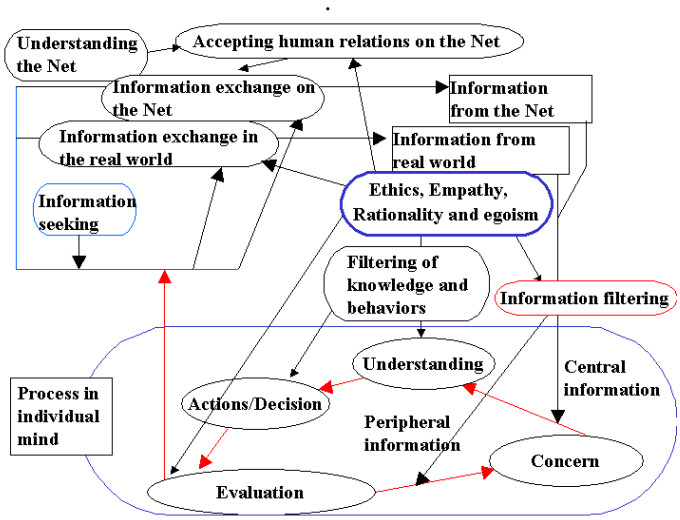

Figure 11 The latest model obtained.

It is noteworthy that the latest model in Figure 11 has similarity to Figure 1 on which our arguments are based, because the target we have been dealing with was the emerging behaviors of people with the awareness of chances in dynamic social environment.

<The New Questionnaire Survey>

For validating the latest hypotheses as in Figure 11 as a process model of chance discovery, we organized a new social survey on peoples' decision process in the real world and on the Internet. Corresponding to each arrow in Figure 11, we made 400 questions in total, for 1007 subjects in US and 1114 in Japan. The samples here were chosen randomly from the panel data of Nihon Research Center, which had been made as a random set of samples from all over Japan and US (excluding the states of Hawaii and Alaska). We made random postal mails for Japan and random digital dialing (RDD) for US for these samples respectively, because the posting system is not equally reliable all over the country in US.

[The Results for US people]

As a result, the result of *KeyGraph* to the answers data for US people was obtained as in Figure 12. This result illustrates a clear cyclic process of chance discovery, validated by the co-occurrence computation among answers. That is, each link and the series of links depict a certain set of people in a specific state, and the connection of these links show the sequence of continuous shifts of the mental contexts of human. Along this context-shift, Figure 12 means US people shifts in the order as:

(1) Concern with peripheral information
(2) Investigation of central information

(3) Perception of risk-factors, first of human (violation of net ethics) and then of nature/society (cancer, traffic accidents, etc).
(4) Perception of and communications for chances, e.g., of buying rare clothes, with strong desire for information.
(5) Weakened desire for information
(6) Evaluation of WWW, and back to new concerns as (1).

Accordingly, we obtained Figure 13 as the validated model of chance discovery, validated in that where all links are abstracts of edges of the *KeyGraph* in Figure 12.

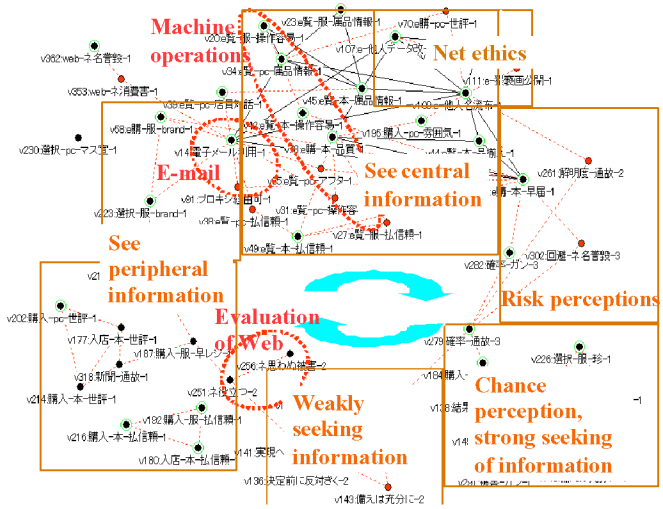

Figure 12. The result of KeyGraph, for US people

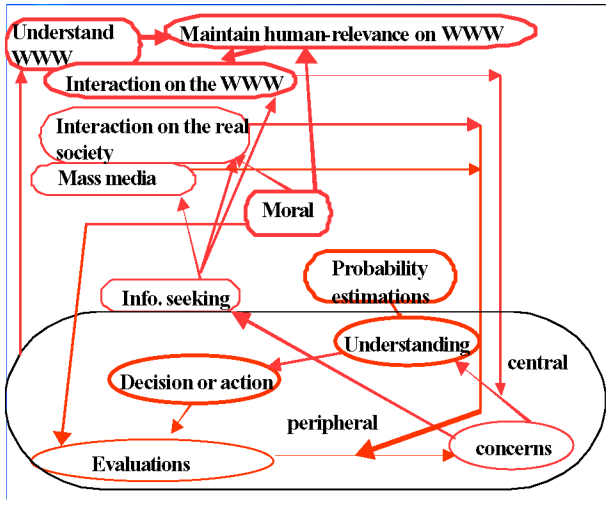

Figure 13 The final model of decision process for the people in US.

[The Results for Japanese people]
On the other hand, the result of *KeyGraph* to the answers from Japanese people was as in Figure 14. This result is composed of the negative attitudes toward the Internet and new information in general. In comparison with the results of subjects in US, the following elements are missed here:

- The cyclic process
- Concern for new information (information-seeking tendencies)
- Morals and risk perceptions as essential parts of the cycles of mental states.
- Understanding and evaluation of their own decisions.

Figure 14. The result of KeyGraph, for Japanese people

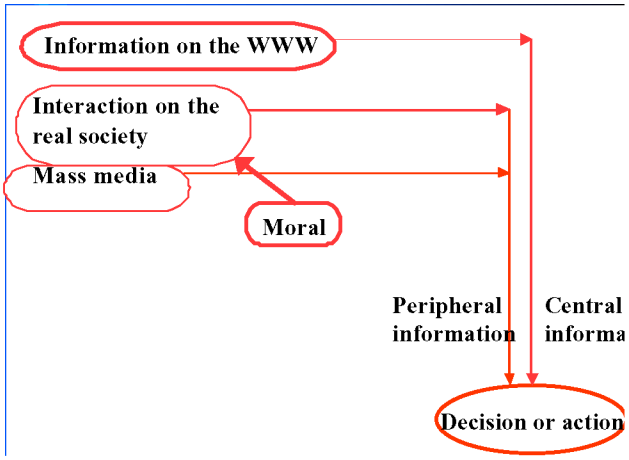

Figure 15 The final model of decisions for Japanese people.

Even though authors are Japanese, this result could not be interpreted as implying our orientation for chance discovery. Reflecting this result, we obtained Figure 15 as the validated model based on the KeyGraph Figure 14 for Japanese subjects.

Conclusion

The double helix model, for the process in which human discovers what we call chances, is exemplified by a case of social survey. The insight to deep-level factors of peoples' behaviors in the real and the virtual worlds are achieved, and the proposals based on the new awareness of rare but significant parts of data were acceptable.

Finally, the double helix lead us to a valid model of the human process of decision making on the information from the real world, mass media, and from the Internet. As a result, the tendencies of US people had a strong orientation to chance discovery, and matched with the cyclic model we had made as the process model of chance discovery. In fact, the fundamental strategies for dealing with significant low-probability changes/events (risks and opportunities) has been developed in US as in the literatures (Ottoo, 2000; Herzenberg et al, 2000; McGrath, R., and MacMillan, 2000), where as Japanese literatures on chances are mostly dedicated to specific domains, e.g., infant education, e-learning, sales-force tasks etc.

The lessons from our current results in the final model enable us to show Japanese people what they are lacking in for overcoming the long and fatal depression of economy. For future plan, we are running the double-helical processes for other areas as medical science, biology, seismology, and marketing.

Acknowledgment: We appreciate the discussants who joined the group interviews for subject data. Chance discovery has various sides to approach, each being supported by the research fund of Japan Science and Technology Corporation (JST), and Scientific Research on Priority Areas (Discovery Science and Active Mining).

References

Simon, H, 1945 *Administrative behavior: a study of decision-making processes in administrative organizations* (1st edition 1945, 4th edition 1997, Simon & Schuster Inc.)

Fukuda, H. and Ohsawa, Y., 2001. Discovery of Rare Essential Foods by Community Navigation with KeyGraph – An Introduction to Data-based Community Marketing, *in Proc. KES2001* (2001)

Ohsawa, Y., 2002. Chance Discoveries for Making Decisions in Complex Real World, *New Generation Computing*, Vol.20 No.2 (2002)

Fayyad, U., Shapiro, G.P. and Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases, *AI magazine*, Vol.17, No.3, 37--54 (1996)

Festinger, L., 1957. A Theory of Cognitive Dissonance, Evanson, IL: Row, Peterson (1957)

Nara, Y., 2001. *Abstracted* in Campus Internet Information Ethics in Japan: An Empirical Study, in *Proc. the 13th Annual Meeting of Western Decision Sciences Institute*, Vancouver, 487.

Ohsawa, et al., 2001 Data-Analysis Competition, Marketing Engineering Section, The Operations Research Society of Japan.

Ohsawa, Y. and Yachida, M., 1999. Discover Risky Active Faults by Indexing an Earthquake Sequence, in Proc. International Conference on Discovery Science.

Matsumura, N., Ohsawa, Y., and Ishizuka, M., 2002., Chance Discovery on Visualizing Correlation Rules, *The 47th SIGFAI workshop, Japanese Soc. AI* (in Japanese)

Ohsawa, Y. and Nara, Y., 2002. Discovery of Virtual Behaviors as Signs of Real Behaviors, *in Proc. KES2001*

Ottoo, R.E., Valuation of Corporate Growth Opportunities: A Real Options Approach (Financial Sector of the American Economy), Garland Publishing 2000

Herzenberg , S.A., Alic, J.A., and Wial, H., New Rules for a New Economy : Employment and Opportunity in Postindustrial America, Cornell University Press (2000)

McGrath, R., and MacMillan, I., The Entrepreneurial Mindset: Strategies for Continuously creating opportunities in an age of uncertainty, Harvard Business School Press (2000)