Polaris: An Integrated Data Miner for Chance Discovery

Naoaki Okazaki[1, 2] *and Yukio Ohsawa*[1, 3]

[1] PRESTO, Japan Science and Technology Corporation
2-2-11 Tsutsujigaoka, Miyagino-ku, Sendai, Miyagi, 983-0852, Japan

[2] Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

[3] Graduate School of Business Science, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012, Japan

e-mail: okazaki@miv.t.u-tokyo.ac.jp

Abstract

KeyGraph, which is an algorithm as well as a tool for discovering rare or novel events, achieves successful outcomes in keyword extraction, earthquake prediction, genome analysis, sales promotion and marketing, questionnaire analysis, and so on. Through such case studies, Ohsawa proposed a double helical model of chance discovery in which humans and data-mining tools co-work; each progresses spirally toward creative reconstruction of ideas. However, existing data-mining tools do not connote architecture to promote chance discovery on the double helix model. We design Polaris, a new data mining system, that features graph representation of a source data as if a user observed a constellation of the data. Polaris accelerates a process of chance discovery by two strategies: it saves work and time for users to be close to their goal, i.e., what they want from the data; and it supports users to convince what they are actually thinking of, providing a way of analyzing their comments for an obtained graph.

Keywords

Human-computer interaction, Chance Discovery, Data miner, KeyGraph

1 Double helical model of Chance Discovery

Studies on Chance Discovery [1] arose with an interest of how computers can help us convince chances, i.e., events significant for decision-making. A large number of data mining tools was proposed or applied in response to the interest. For instance, KeyGraph [2], which was originally invented as a keyword extraction algorithm, finds rare or novel events from basket data and achieves successful outcomes in earthquake prediction, genome analysis, sales promotion and marketing, questionnaire analysis, etc.

Data mining tools may indeed help us dig up chances, but be nothing more than a tool; computers can show only an objective report on the data. We must interact with computers tightly and train ourselves strictly because it is us that find chances. We often tend to rely on the tools so much that we forget to prepare our mind for chances.

Ohsawa proposed a double helical model of chance discovery [3]. He distinguished a spiral process of human awareness of new chances and a process of computers that receive, mine and visualize data. As each helix progress spirally, interaction with each helix promotes creative reconstruction of ideas (Fig. 1).

Let us take questionnaires for example. We draw up and send out a questionnaire since we have concerned with people's impressions for some targets. After we collect responses to the questionnaire, we make a data that can be passed to a data-mining tool. We try to understand the visualized data obtained from the data-mining tool and record thoughts of human as a text to discover chances. A text-mining tool is applied to the texts to externalise subject's concerns with hidden factors. This process to look at one's own mind is not covered in double-helix model. That is, the name "double-helix" means the parallel processing, i.e., the simultaneous runs of this pair of helixes (spiral processes), due to the input if "the subject data" that

monitors the mind of the subject who tries to discover chances. In previous data mining, the computer was taking a rest while the subject was discussion the last output of "the object data" that is the data for the target problem.

Between these two helixes, interactions occur: a subject-data is obtained in 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 results from the subject-data (DM-a, in Fig. 1) are reflected on the subject's understanding of his or her own concern, and the mining results from the object data are reflected on the understanding of the chance. In the remainder, we show the double-helix model as a new method for hypothesis creation, exemplified by a sociological example.

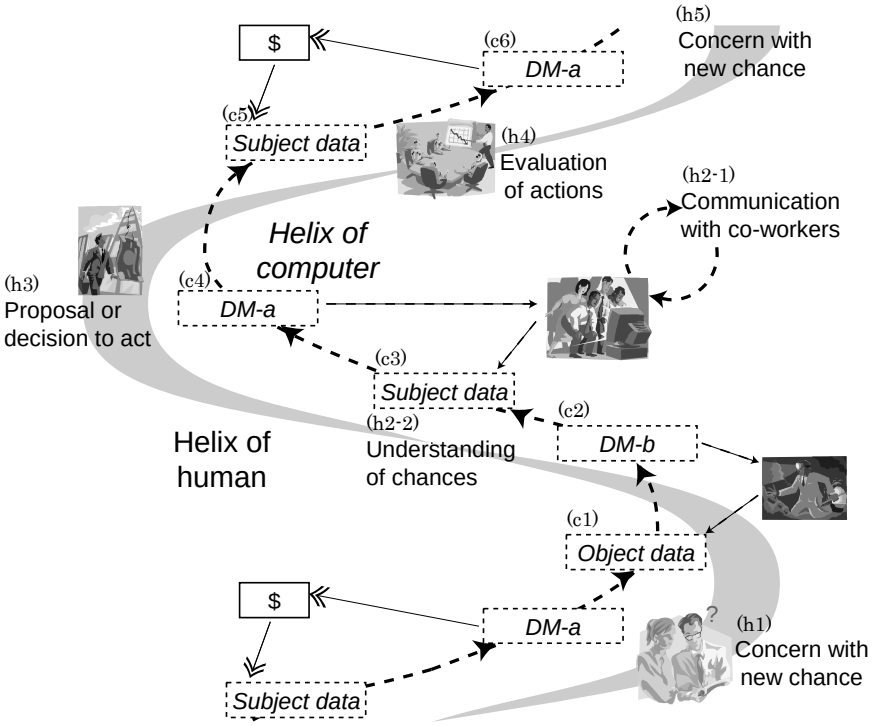

Figure 1: The double helical process of chance discovery.

It is preferable that computers and we advance the spiral processes quickly and smoothly. Speed is important factor when we use in competitive society, i.e., business applications. Additionally, we must walk a rocky road to discover a chance as implied by the phrase "throes of creation". If computers cannot shed light on the road, we may find it very hard to put ourselves ahead; as a worst case, we may lose our ambition for chances.

Let us consider how computers accelerate process of chance discovery in the double helical model. You may hit upon an idea of accelerating data mining itself, e.g., optimising KeyGraph algorithm and source code for speed. However, we can obtain an output from KeyGraph typically within ten seconds. That does not matter so much; we must take notice to other points to impact overall speed of the processes.

It goes without saying that acquisition of object-data is to collect data we want to inquire. However, it implies another task we must perform: to convert the raw data into a worked data that can be passed to the data-mining tool. Take KeyGraph again for example. KeyGraph requires basket data in the form of text: each item must be delimited by a space; and each basket must be delimited by a period. It is essential in advance to make such a basket data that we decide a subset of the data we are going to analyse and unit of items and baskets (cooccurrences). This task sometimes requires special tools or programming of text processing because it is hard that we perform this task in a text editor. Existing data miners do not take care much of this important task that influences data-mining result although we usually perform this task more than once until we hit our goal.

After we concern about and understand the output of data mining, we express ourselves in writing and make a "subject data", e.g., writing our thought, discussing with colleagues about the resultant output or setting up hypotheses. We analyse the "subject data" by a text miner to clarify what we have in mind and

validate / invalidate the hypotheses. However, no existing data miner takes this phase seriously; we used to make a text for analysing our thought by hand.
KeyGraph as a data-mining tool convinced us double helical model of chance discovery. Now it is time a data-mining tool evolved to connote chance discovering process on the double helix model. We design a new data-mining framework for chance discovery in this motivation.

2 Polaris

We propose Polaris, a new data-mining framework derived from common processes of chance discovery based on the double helix model. Polaris outputs graph representation of a source data as if the user observed a constellation of the data. Polaris is not a term that points a specific data-mining method but affords a platform for users to interact with data-mining tools. Although we integrate KeyGraph algorithm into Polaris in our first attempt, we can easily integrate other data-mining method that facilitates graph representation of a source because we separate data-mining component from Polaris.
Polaris has two major features to promote entire processes for chance discovery. Firstly, Polaris saves work and time of users to obtain what they want from the data or bring the users close to their goal. For example, we often encounter a graph such as Fig.2 (a), in which a node connected to many other nodes in the graph. Node “A” is actually important because the node co-occurs with many other nodes. However, it wields influence over the nodes so strongly that we cannot catch how other nodes are connected. In this case, we usually add node “A” to stop word list to eliminate node “A” from the source data; Fig. 2 (b) may be more comprehensible to us.

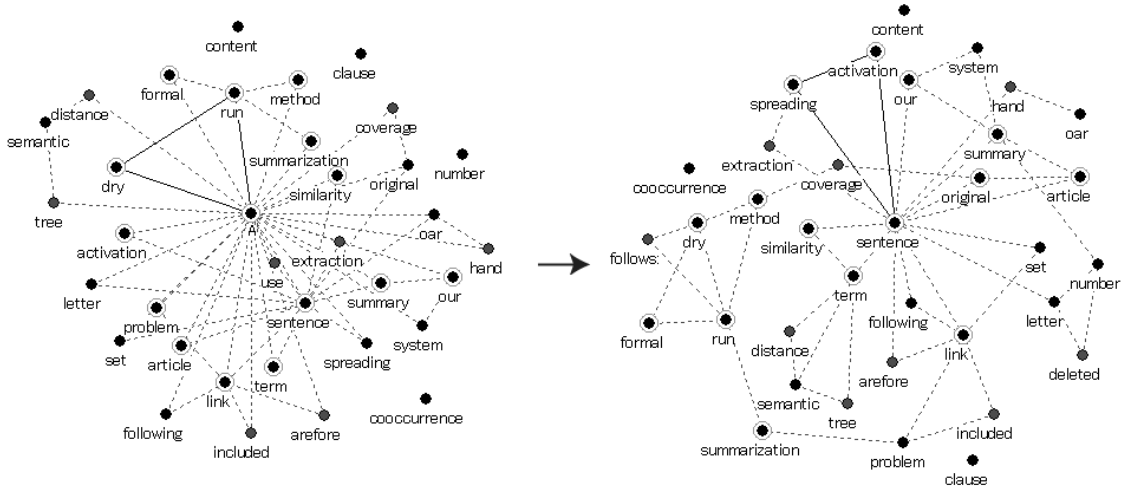

Figure 2: Pre-processing of raw data. We should eliminate node ‘A’ from the worked data.

We described in the previous section that we usually repeated such pre-processing and data mining. Polaris supports typical kinds of pre-processing as follows.

- Morphological analysis (used for Japanese text mining)
- Stop list management (we can add a node to stop list by right-clicking of the node)
- Node dictionary management (we can equate a node with another)
- Basket extraction filter (we can extract baskets that matches a rule and reanalyse them)

The second major feature of Polaris is to supports users to convince what they are actually thinking of. We indicated the use of subject data for validating our understandings or hypotheses in the previous section. Polaris has a built-in comment editor where we take minutes of discussion or chat; we can analyse the comments or hypothesises by selecting a menu item.
Fig. 3 is a screenshot of Polaris. We can see four sub-windows (data view window, file list window, additional information window and comment window) in this figure. After Polaris receives a source data and work upon the data for a data-mining tool, it passes the worked data to the data-mining tool; it visualizes an obtained graph in the data-view window. Polaris computes optimal arrangement of nodes

automatically by spring model [4]; we can rearrange nodes by drag-and-drop operation according to our own interpretation of the graph. When we click on a node with the right mouse button, we can configure pre-processing by selecting a pop-up menu (e.g., we can add the node into the current stop list or extract baskets that contains the node).
Additional information window shows supplementary information about a selected node, e.g., frequency of the node, list of cooccurrence with the node, the source data where the node occurs (see Fig. 3), etc. Additional information of nodes serves an intermediary between source data and mined data; we deepen our understanding of why we obtained this graph.
Polaris supports users to convince what they are actually thinking of, providing a way of analysing their comments or hypothesises for an obtained graph. We can put together in file list window multiple relevant data files and our findings written in comment window.

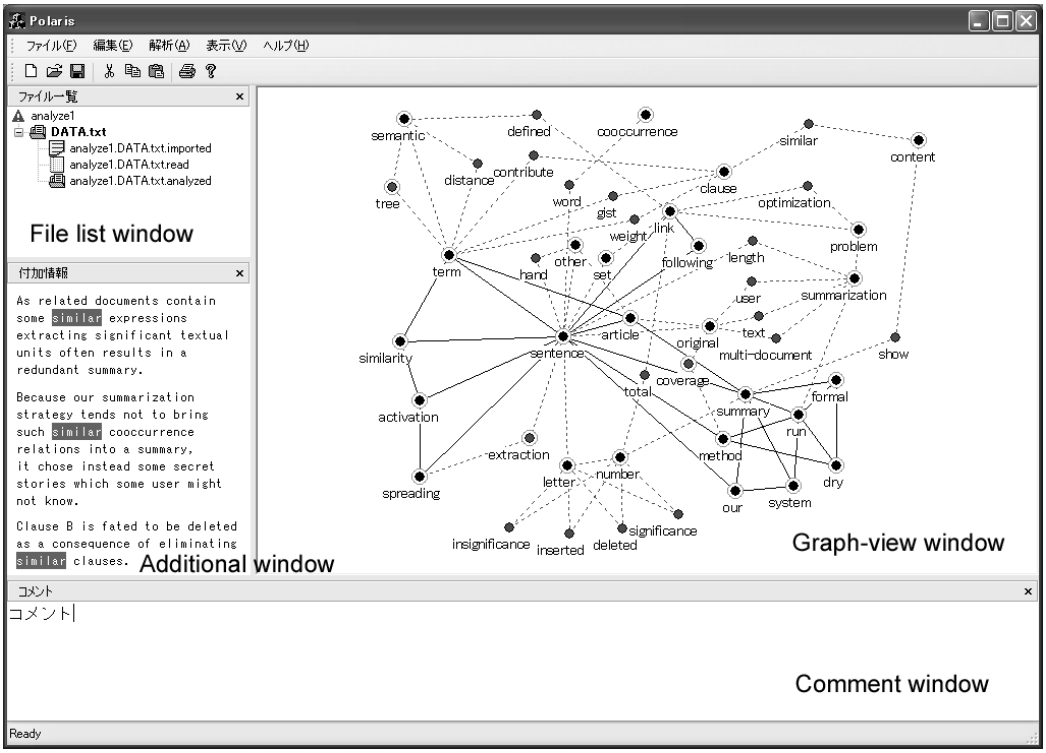

Figure 3: A screen shot of Polaris (under development).

3 Conclusion

We show Polaris, an integrated data miner for Chance Discovery. It is aimed at providing a new framework that promotes double helical processes. Although Polaris is under development at this moment, we are looking forward to see how it has an impact on chance discovery.

References

[1] Yukio Ohsawa (2002). Chance discoveries for making decisions in complex real world. *New Generation Computing* (Springer-Verlag and Ohmsha, Ltd.), Vol. 20, No. 3, pp.143-163.

[2] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida (1998). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor, In Proc. of Advanced Digital Library Conference (IEEE ADL'98), pp.12-18.

[3] Yukio Ohsawa and Yumiko Nara (2003). Decision process modelling across Internet and real world by double helical model of chance discovery. New Generation Computing (Springer-Verlag and Ohmsha, Ltd.), Vol.21 No.2, pp.109-122

[4] Peter Eades (1984). A heuristic for graph drawing. *Congressus Numerantium,* Vol. 42, pp. 149-160.