チャンス発見のための統合型データマイニングツール Polaris

Polaris: An Integrated Data Miner for Chance Discovery

岡崎　直観 [1]　　大澤　幸生 [2]　　石塚　満 [1]
Naoaki Okazaki　　Yukio Ohsawa　　Mitsuru Ishizuka

[1] 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo

[2] 筑波大学大学院経営システム科学専攻
Graduate School of Systems Management, University of Tsukuba

Abstract: Chance Discovery, which is a new field of research to find some events or situations with significant impact on human decision-making, differs from existing Data Mining in that Chance Discovery takes particular note of rare events and human-computer interaction to convince us the significance of the events. A great deal of data mining methods was proposed or applied in response to the central interest of how computers can help us convince chances. 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. 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, the double helical model takes a good amount of time because it requires state transitions of a human's mind. We propose Polaris, a new data-mining framework, to promote the double helix process of chance discovery by:

- Reducing the cost of cleansing a textual data for mining;
- Supporting human's insight and understanding of a visualized data.

Although Polaris was originally an extension of KeyGraph to connote the double helix process of chance discovery, we can integrate other data mining methods that take a document matrix as an input. It consists of three components: "text reader component" to convert a text data into a document matrix; "mining component" to analyze the document matrix; and "visualization component" to visualize the analyzed data.

1. はじめに

近年，企業間の競争が激化していく中，電子化文書やインターネットなどの情報技術を活用し，コールセンターの問い合わせ内容, アンケート調査, 企業が主体となって運営するウェブ上のコミュニティ等を利用し，顧客要求（VOC: Voice Of Customer）を収集する企業が増えている. 蓄積したデータをうまく活用すれば，顧客の期待に応える意思決定を行うことができるが，データの量が膨大になると，顧客要求を整理することさえ困難となる．このような大量のデータからビジネスに活用できる知識を獲得するための技術としてデータマイニング [1] が注目され，様々な研究・製品が発表されている．また，企業に蓄積されているデータの約 80%は構造化されていないテキストであると言われており，テキストデータから興味深く重要なパターンを抽出し，その分析結果を視覚化するテキストマイニング [2] [3] も注目を浴びている．

一方，人の意思決定において重要な事象を発

岡崎　直観　東京大学大学院情報理工学系研究科
〒113－8656　東京都文京区本郷 7-3-1
Tel: 03-5841-6755, Fax: 03-5841-8570,
E-mail: okazaki@miv.t.u-tokyo.ac.jp

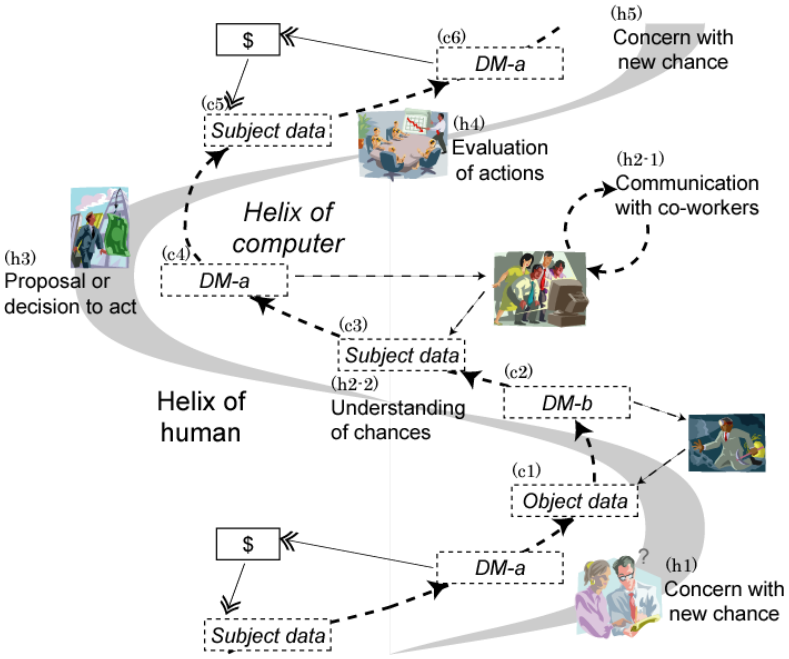


図１．チャンス発見の二重螺旋プロセス

見するための新しい研究領域として，チャンス発見［4］が開拓された．チャンス発見と既存のデータマイニング技術の大きな違いとしては，チャンス発見が希少な事象を重視している点，コンピュータが提示した事象を人間が理解して行動に移すための人間とコンピュータのインタラクションを重視している点が挙げられる．

手元に稀な事象を重視するデータマイニング・アルゴリズム（例えば KeyGraph [5] など）があるとしよう．そのアルゴリズムを活用して知識マネジメントやチャンス発見のプロセスを円滑に進めるには，視覚化された分析結果に対する理解と，その分析結果から想像されるシナリオ（仮説ルール）の発案・深化を促進する必要がある．このタスクは人間の視覚化データに対する理解力・想像力が主体となって進められるものであるが，このタスクの成否がチャンス発見のプロセス全体の成否を左右すると言っても過言ではない．ゆえに，そのタスクにふさわしいインターフェースを考案し，人間の想像力を引き出すような環境を整えることは，チャンス発見システムの中心的課題と言える．本論文は，チャンス発見の二重螺旋モデルを簡潔に導入した後，チャンス発見のための新しいデータマイニング・フレームワークである Polaris というシステムを提案するものである．

2. チャンス発見の二重螺旋モデル

チャンス発見の中心的なテーマとして，人間がチャンス（意思決定のために重要な事象）に気づくために，コンピュータの側からどのような支援を行えばよいのかという課題がある．この問いに対して，数多くのデータマイニング・ツールが適用・提案されてきた．しかしながら，データマイニング・ツールは我々がチャンスを掘り起こす手助けはできたとしても，与えられたデータに関する客観的な報告を行うものであり，やはり人間が使う道具の域を出ない．チャンス発見においては，ユーザーが自分の潜在的な目的に気づいていないことがほとんどであるので，人間の果たす役割が重要になる．

大澤［6］はチャンス発見のための二重螺旋モデル（図１）を提案し，人間が新しいチャンスに気づく螺旋状のプロセスと，コンピュータがデータを受け取ってマイニング・視覚化するプロセスを分離し，それぞれの螺旋プロセスが並進しつつ，インタラクションが起こることによって新しいアイディアが形成されていくと唱えた．二重螺旋モデルでは，人間のチャンス発見の螺旋プロセス（「関心→理解→発案→行動」という繰り返し）と並行し，その人間が関心を寄せる解析対象から観測されるデータと，その人間の思考内容のテキストデータに対して，コ

ンピュータがマイニングを繰り返す.

この二重螺旋モデルの実践にあたっては，人間が心理状態の遷移を順番に経ていくので，発見までの道のりに時間がかかるという問題点が指摘されている．人間とコンピュータは，この螺旋プロセスをスムースに進むのが理想的である．なぜなら，スピードや作業コストはビジネスなどの競争社会の中では重要なファクターであるし,「生みの苦しみ」と表現されるように，我々がチャンスを発見するまでの道のりは決して平坦ではないからである．チャンスを発見するまでの道のりがあまりにも大変だと，解析対象データへの関心が薄れるし，新たな知見を見出すことそのものへの興味も減少してしまう.

このチャンス発見のプロセスを加速するにはどのようにすればよいだろうか？ まず,データマイニングの解析速度を向上させることを思い浮かべるかも知れないが, 例えば KeyGraph が解析に要する時間は数十秒程度のオーダーであるので，このフェーズを改善したとしてもプロセス全体に与えるインパクトは少ない．マイニング結果から得られた仮説を実環境で適用し，その仮説を評価するフェーズは，システムが支援できる範疇を超えている.

しかし，収集したデータを初めてマイニングする場合や，マイニング結果から得た仮説をすでに蓄積したデータから検証する場合，そのデータのクレンジングに要するコストを下げることは大きな意味がある．なぜなら，チャンス発見のプロセスの一翼を担うのは，データマイニングの専門家ではなく，実環境に日頃から接している関係者だからである．このフェーズで躓いてしまうと，自分の目的を掘り起こしたり，その目的を満たす視覚化データを得ることが困難となる.

二重螺旋プロセスにはマイニング結果に関心を持ったら，その結果に対する解釈を他の仲間と議論しつつ，仮説ルールを文章で生成するというフェーズがある（図１の h2-1, h2-2）．発案された仮説ルールを解析対象データとしてテキストマイニングを適用することで，仮説の洗練や観察されている現象への理解を深めるのであるが，このフェーズをテキストマイニング・システムの中に組み込むと，我々が考えたシナリオを即座に視覚化することが可能となり，チャンス発見プロセス全体の加速に貢献する.

3. Polaris

3.1. Polaris とは

Polaris[1]は，チャンス発見の二重螺旋プロセスをもとに新たに設計されたデータマイニング・フレームワークであり，今まで述べてきたようなアプローチを使って，KeyGraph をチャンス発見プロセスに適合するように拡張したシステムと捉えると分かりやすい．チャンス発見の中心的役割を担ってきた KeyGraph は，もともとキーワード抽出のためのアルゴリズムであり，チャンス発見における KeyGraph の利用形態をヒントに二重螺旋モデルが生み出された．KeyGraph は二重螺旋プロセスの中でデータの解析を担ってきたが，チャンス発見の中心的課題と言えるデータマイニング・ツールと人間とのインタラクションを支援するという面においては，不十分な感があった．このような問題意識のもと，二重螺旋プロセス全体を支援するシステムとして Polaris が生まることとなったが, 単に KeyGraph を拡張したテキストマイニング・システムではなく，テキストデータを受け取って何らかの解析処理を行って視覚化するツールとして一般化している．現在，Polaris には解析手法として KeyGraph が実装されているが, 将来的には KeyGraph 以外の解析処理も実装される予定である.

解析したいテキストファイルを指定すると，形態素解析，不要語除去などの読み込み処理から，KeyGraph 等によるマイニング，データの視覚化まで，すべて Polaris 上で実行できる．多彩な読み込み処理に加えて，解析結果と元テキストデータとの照らし合わせ，解析フォーカスの設定などの機能により，データマイニングを適用するまでの前処理とマイニング結果の理解を支援し，ユーザーを見つけたい現象へ素早く近づけることが可能である．また，視覚化データを見て思考した内容を書きとめておくウィ

[1] 北の夜空には，カシオペア座やおおぐま座（北斗七星）があり, それらの星の位置関係を頼りに北極星を見つけ出すことができる．北極星は二等星の比較的明るい星であるが，周辺にあるほかの星との関係を参照することで格段に見つけやすくなる．我々がこれから提案するシステムは，情報をノードとエッジで構成されるグラフ構造で視覚化し, その関係を眺望することによって重要な事象（ノード，もしくは星）やそのノードを見出した背景（ノード間の関係，もしくは星座）を見出すものである．このようなアナロジーにより，我々は本システムを Polaris（英語で北極星の意）と命名した.

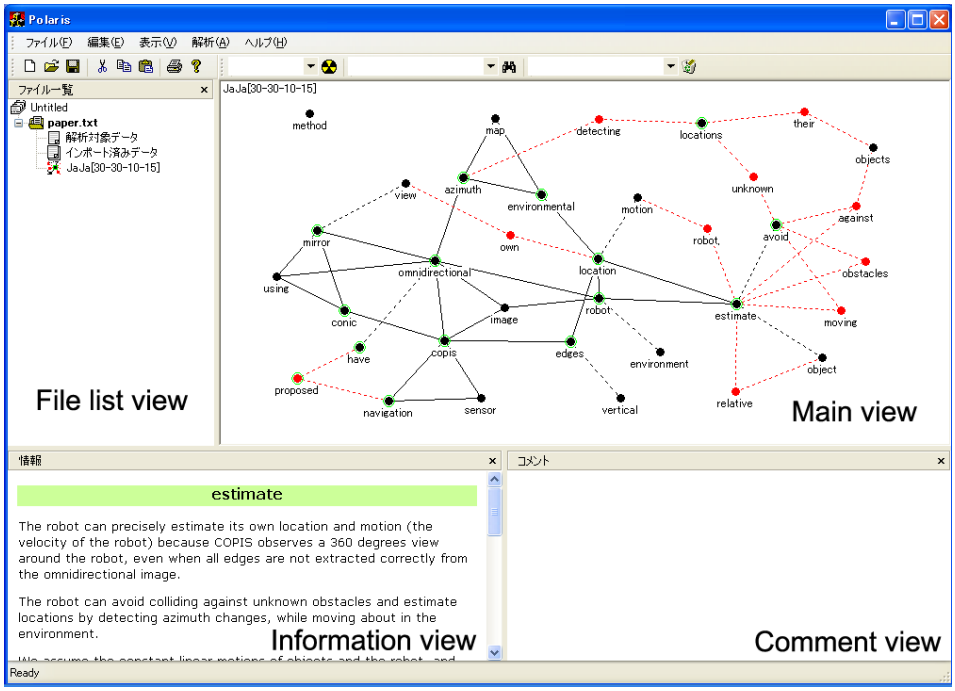


図２．Polaris で英語のテキストを解析している際のスクリーンショット

ンドウが用意されており，解析結果から得られた仮説を文章形式で書きとめ，その文章自体をマイニングしながら，発案した仮説を深化・検証することができる．

図２は英語のテキストデータを Polaris に与え，KeyGraph でマイニングして視覚化した際のスクリーンショットである．Polaris は解析結果を表示する「Main view」，現在開いているウィンドウを列挙する「File list view」，付加的な情報を表示する「Information view」，解析結果を見ながら気づいたことを書き込む「Comment view」の４つのウィンドウから構成されている．

KeyGraph でテキストマイニングをする場合，解析対象テキスト中の語とその関係が「Main view」ウィンドウにノードとエッジとして表現される．この解析結果の解釈を進めるにあたっては，例えばグラフ中に現れた単語が元のテキストで使われて箇所を読み，そのノードが存在する背景を調べるという作業が行われる．今まで，この作業は元のテキストをエディタ等のソフトウェアで開き，検索を行うことで実現していたが，Polaris では文脈を調べたいノードをダブルクリックするだけで，元テキストで使われていた箇所を「Information view」ウィンドウに表示させることができる．

3.2. システム概要

図３に Polaris のシステム概要を示す．解析したいテキストデータが与えられると，「テキスト読み込み」，「マイニング」，「視覚化」コンポーネントを順に経て解析結果が出力される．

テキスト読み込みコンポーネントは，解析対象テキストを読み込み，加工・クレンジングして解析アルゴリズムで扱える文書行列形式に変換する．このコンポーネントは，何種類かの読み込みモジュールを用途に合わせて切り替えることができ，日本語テキスト，英語テキスト，単純なバスケット・データなど，様々なテキストデータ形式を吸収するものである．日本語テキスト読み込みモジュールでは，句読点を利用した文同定，形態素解析，条件付き抽出（ある語を含む文のみを抽出する等），不要語フィルタ（不要語リストに含まれる語は読み込まない），語の表記の統一（類義語辞書による統一，英字小文字への変換など）をサポートしている．

マイニング・コンポーネントは，データ読み込みコンポーネントが作成した文書行列を解析し，その結果を出力する．現在は KeyGraph アルゴリズムのみが用意されているが，文書行列を対象とするマイニング手法であれば実装可能であり，将来的には複数のマイニング手法を用途に応じて選択できるようにする予定である．

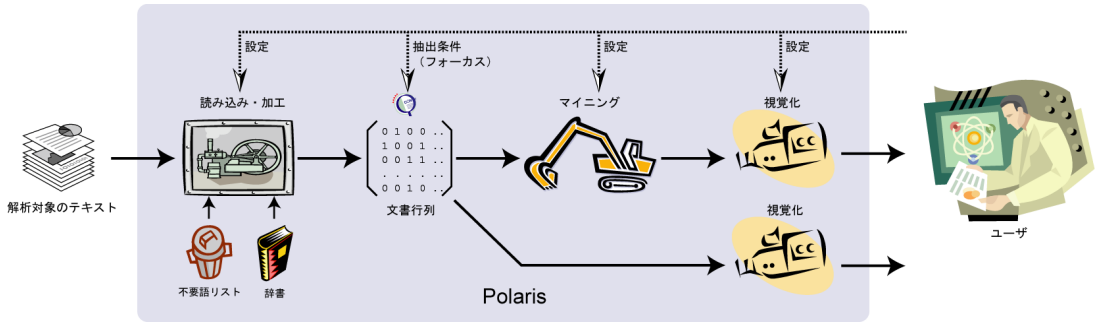


図３．Polaris のシステムアーキテクチャ

視覚化コンポーネントは，データの分析結果や解析対象データそのものを表示する．Polaris では解析対象データに対して，複数の視覚化チャネルを持つことを許容しており，ユーザーはグラフやテキストなど様々な視覚化チャネルから解析データを眺めることができる．グラフ表現の視覚化には，ばねモデル [7] による最適配置が行われる．

3.3. 文書行列に対する操作

テキストに対して KeyGraph を適用したグラフを眺めていると，あるノードの存在が邪魔に思えることがある．図４はその典型例を示したものであるが，図４(a)において”a”というノードが他の多くのノードと結ばれてしまい，その他のノード同士の関係が見づらくなっている．”a”というノードが重要なものであるならば，図４(a)の結果は納得のいくものであるが，この場合の解析対象は英語テキストであり，”a”は冠詞であるから，”a”が他の多くの語と関係を持つのは自明である．そこで，”a”という単語を文書行列から取り除いて再び解析すると，図４(b)の結果が得られ，グラフの情報が改善される．テキストマイニングにおいては，文章を特徴付けない語を，不要語として除去するが，不要語は解析対象のテキストのドメインに依存するので，解析結果を眺めながら不要語リストを調節する必要がある．

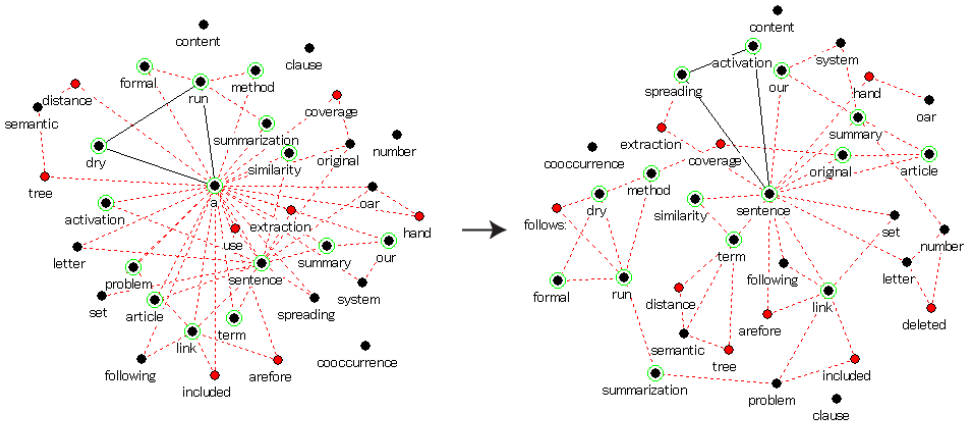
Polaris ではこのような文書行列に対する操作をグラフ表示などの視覚化データの上からマウス操作で行うことができる．ある語を文書行列から消去する操作の他に，ある条件を満たす行／列要素のみを残すという操作が行える．例えば，あるテキストを KeyGraph で解析したグラフを眺めている時，「不満」という語に興味を持ち，このテキストで述べられている不満とはどのようなものなのか知りたいと考えたとする．この場合，「不満」というノードに対して簡単なマウス操作を行うだけで，「不満」という語を含む文のみを解析し，ユーザーが持った興味に対して即座に答えることができる．

4. 関連研究

既存のテキストマイニング・システムは，ある特定の解析アルゴリズムを搭載する「専用ツール」と，複数の解析アルゴリズムを搭載し，解析手順を自由に設計できる「汎用ツール」に大別できる．専用テキストマイニング・ツール（例えば KDT [3], KeyGraph [5]）は，ユーザーの与えたテキストデータと解析のためのパラメータを受け取り，解析結果をユーザーに返すというシンプルな構造をしており，チャンス発見のプロセスにおいて重要な，解析結果を人間が理解し，シナリオを生成するタスクに対する支援は行われない．また，ツールによっては蓄積されていたテキストデータを，解析可能な形式に変換する必要が生じ，データマイニングの専門家ではないチャンス発見の主体者がこれらの専用ツールを使いこなすのは難しい．

汎用テキストマイニング・ツール（例えば SPSS 社の Text Mining for Clementine，IBM 社の Intelligent Miner for Text，TextVis [8]）は，複数の解析アルゴリズムを搭載すると共に，それらの解析手順を自由に設計したり，データのクレンジングを統合環境上で実行することができる．多機能であるため，汎用ツールは使いこなせるようになると便利である．

Polaris は専用ツールと汎用ツールの中間に位置すると言える．Polaris はチャンス発見のための多目的・統合型データマイニング・フレームワークであるが，Polaris を利用する人がデータマイニングの専門家でないことを想定し，



(a) Node 'A' cooccurs with too many nodes.　　(b) Reanalyzing after eliminating node 'A' from the data.

図４．不要語除去の例

解析対象データを「読み込み」→「マイニング」→「視覚化」という手順で必ず処理するように制限している．その代わり，各々のコンポーネントを複数用意しておいてユーザーに選択させ，設定画面を使って各コンポーネントの挙動を制御することで，用途の幅を広げつつ，比較的簡単に使用できるように工夫している．

5. 結論

チャンス発見のための統合型データマイニング・フレームワークとして Polaris を紹介した．現在，Polaris は開発途上のシステムであり，既存のシステムとの比較や評価はまだ行われていないが，今後 Polaris の開発を続け，チャンス発見に与えるインパクトを調べる予定である．

参考文献

[1] Usama M. Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth (1996). From data mining to knowledge discovery in databases, AI magazine, Vol. 17, No. 3, pp.37-54.

[2] Marti A. Hearst (1999). Untangling text data mining. In Proc. of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, *(invited paper)*.

[3] Ronen Feldman and Ido Dagan (1995). Knowledge discovery in textual databases (KDT). In Proc. of the First International Conference on Knowledge Discovery (KDD-95), pp.112-117.

[4] 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.

[5] 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.

[6] 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

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

[8] David Landau, Ronen Feldman, Yonatan Aumann, Moshe Fresko, Yehuda Lindell, Orly Lipshtat, and Oren Zamir (1998). TextVis: An Integrated Visual Environment for Text Mining. In Proc. of the 2nd European Symp. on Principles of Data Mining and Knowledge Discovery (PKDD '98), Lecture Notes in Artif. Intell. 1510, pp.56-64. Springer-Verlag.