解説

ネット上の情報を分析する技術[†]

櫻井 茂明¹ 小山 聡²

1 はじめに

コンピュータ及びネットワーク環境の進展に伴って、情報を発信する人の数が急速に拡大しており、膨大な量の情報が蓄積されるようになった。しかしながら、人間の情報処理能力は限られており、せっかく蓄積した情報が必ずしも有効には活用されていない。このような状況が情報洪水という言葉で語られるようになって久しくなるが、近年この状況は益々加速しており、このような状況を打破する技術として、統計学や機械学習技術に基づいた各種手法が開発されている。当初は、比較的単純に構造化された数値や離散値を対象とした技術が研究対象として取り上げられていたが、より複雑に絡み合った情報を分析したいといったニーズが高まっており、テキストデータ、グラフ構造を持ったデータ、XML データに代表される半構造データ、時系列データを分析する技術へとその研究対象がシフトしつつある。これらの技術がインターネット上に散逸するデータを分析する技術として注目を集めており、以下の章ではこれらのデータを分析する技術に焦点をあてて解説を行う。

本解説では、2章で従来から研究されているデータ分析技術を簡単に紹介し、3章でテキストデータを分析する技術を紹介し、4章で Web 構造を分析する技術を紹介し、5章でマイニング技術を応用した例を紹介し、6章でまとめと最終的な今後の展望を述べる。

2 データを分析する技術

数値や離散値からなるデータを分析する技術としては、統計学が古くから研究されており、数値データ間の関係をモデル化する回帰分析や、離散値間の関係をモデル化する数量化Ⅲ類が広く利用されている。

また、統計学をそのベースとしつつも、人工知能分

野における考え方を考慮した機械学習法の研究も進んでいる。

機械学習法を大きくふたつに分けると、教師を利用しないタイプと教師を利用するタイプに分けることができる。教師を利用しないタイプとしては、クラスタリングや Apriori アルゴリズム [1] が知られている。クラスタリングは、与えられたデータ集合を、データの特徴付ける何らかの特徴量に従って複数の部分集合 (クラスター) に分割する方法であり、初期に提案された K means 法が有名である。K means 法は、予め指定する K 個のクラスターにデータ集合を分割する方法であり、クラスターの中心として最初に選択するデータに、生成されるクラスターが依存することが知られている。K means 法以外にも、多数のクラスタリングが提案されており、階層的なクラスターを生成するものの、各データがひとつのクラスターに所属するのではなく複数のクラスターに所属するもの等が提案されている [2]。これらのクラスタリングは、ある観点で類似するデータをグループ化することができ、利用者に関連深いデータを提示することができる。このため、利用者に何らかの気づきを誘導することができる。

一方、Apriori アルゴリズムは、主にバスケット分析において利用される方法である。ひとつのバスケットに複数のアイテムが含まれるとの前提の下、同時に購入される頻度の高いアイテムの組み合わせを発見する。Apriori アルゴリズムは、同時に購入されるアイテムの数を増やしていくことにより、頻度の高いすべてのアイテム集合を発見する。また、支持度と確信度に基づいて、あるアイテム集合が選択された場合に、同時に選択されやすいアイテムを識別する相関ルールを意味のあるルールとして発見する。ここで、支持度は、全バスケットに対してルールの条件部と結論部に含まれるすべてのアイテムが同時に選択される回数の割合として定義されており、確信度は、条件部のアイテムが同時に選択された回数に対して条件部と結論部に含まれるすべてのアイテムが同時に選択された回数に対する割合として定義される。Apriori アルゴリズムにおいては、頻度が高いアイテムや意味のあるルールの識別を、しきい値に基づいて判断しているため、適切なし

[†] Techniques for Analyzing Information on the Net
Shigeaki SAKURAI, Satoshi Oyama

1 ㈱東芝 研究開発センター

Corporate Research & Development Center, Toshiba Corporation

2 京都大学

Kyoto University

きい値を設定しないと、意味のないアイテムの組み合わせやルールを多数発見し、その発見に多大なる時間を必要とする。このため、意味のあるルールを測る各種指標が提案されている [33]。

教師を利用するタイプでは、データに対して教師から与えられるラベルを利用することにより、データとそのラベルとの間の関係を学習したり、その関係に基づいて新たに与えられたデータの分類を行ったりすることができる。このタイプの学習法としては、データとその結果との関係を木構造で表現した決定木を学習する ID3 [27] が有名である。ID3は、データの特徴付けの属性を用いてデータを分割した場合に得られる相互情報量に基づいて、データを分割する属性を決定し、データ集合を部分集合に分割する。この部分集合に対して、再帰的に同様な分割を繰り返すことにより、決定木を成長させていく。当初提案された ID3は、離散値からなるデータを対象としていたが、数値やファジィ値からなるデータ [29] [35] や、上位-下位のような構造を持ったデータ [31] をも扱えるように拡張がなされている。

また、データとそのラベルとの関係を超平面で表現した SVM (Support Vector Machine) [36] も知られている。SVM は、与えられたデータ集合を高次元空間に写像することにより、教師から与えられたラベルを正しく分離する超平面を決定する。このとき、予め指定するカーネル関数を利用して写像を行うことにより、高速に超平面を決定することができる。SVM は、二次元最適化問題として定式化でき、数学的なバックグラウンドも整備されているため、近年活発に研究がなされている。また、サンプルデータに対する解析解を求めることにより、メモリ容量や計算量を大幅に削減する方法も提案されている [26]。

Web ページの分類では、ページ間のリンク関係も重要な情報になるものの、ID3や SVM 等のアルゴリズムでは、このようなオブジェクト間の関係を利用した学習を行うことは困難である。この問題に対して、一階述語論理に基づいた帰納論理プログラミング [28] は、オブジェクト間の関係を利用した学習を行うことができる。すなわち、“Web ページ A から B へリンクがある”といった条件を述語で表すことにより、分類規則の中で利用することができる。このようなオブジェクトの関係を扱う学習は、関係学習 (relational learning) と呼ばれており、近年盛んに研究がなされている。また、確率モデルと組み合わされた relational probability model についての研究も行われている [25]。

教師無し学習と教師あり学習の中間に位置する学習方法として、統計的手法に基づく方法が知られている。

Naive Bayes Classifier [19] は教師を用いた学習法であるが、クラスタリングに用いられる EM アルゴリズムと組み合わせることにより、少数のラベル付けされたデータと、大量のラベル付けされていないデータから、精度の良い学習を行うことができる [20]。これは、ラベル付けされていないデータであっても、大量に存在する場合には、その統計的な性質を分類に利用できるからであり、ラベル付けされていないデータ (ページ) を比較的容易に収集できる Web ページの分析にとっては、有望な学習法となっている。

この他の機械学習法としては、ニューラルネットワークが知られている。ニューラルネットワークは、そのタイプにより、教師を利用しないものと教師を利用するものが提案されている。前者のタイプとしては、二層構造のネットワークの特殊なものである SOM (Self-Organizing Maps) [15] が知られている。SOM は、入力データに対して勝者となるニューロンを、その周囲のニューロンとともに強化する。このため、入力データ集合に内在するトポロジーをニューロン群のトポロジーと一致させるように学習させることができる。学習されるマップは通常二次元上に入力データを配置しており、高次元の特徴量で特徴付けられるデータ間の関係を視覚的に把握することができる。一方、後者のタイプとしては、階層型のものが頻繁に利用されている。学習方式としては、バックプロパゲーションが有名であり、出力層から与えられた結果と教師から与えられた結果から誤差を計算し、その誤差に基づいて前の階層における重みを調整する。このように、ある階層における誤差に基づいて、前の階層の重みを調整していくことにより、入力データとその結果との間の関係を学習する。このようなニューラルネットワークは、入力データとその結果との間に内在する非線形的な関係を表現することができる。

以上に紹介した多くの技術はツール化も進んでおり、利用者は GUI を通して簡便にデータを分析することができる。また、これらのツールは、ネットワークからデータを収集するシステムやデータベースシステムとの連携も深めている。しかしながら、与えられたデータに対して、どのような分析を実施し、その中からどのような知見を得るかは、利用者のセンスに依存するところであり、分析ツールを導入したからといって、データから有効な知見が必ずしも得られる訳ではない。

3 テキストデータの分析

3.1 テキストマイニングとは

1990年代後半からテキストデータを分析する技術として、テキストマイニングという用語が頻繁に用いら

れるようになった。テキストマイニングを従来のデータマイニングと対比して見れば、データマイニングが構造化されたデータを対象とするのに対して、テキストマイニングは構造化されていないテキストデータを対象とする。テキストマイニングは、データマイニング技術と自然言語処理技術の境界領域上の技術として実現されるため、各サイドの国際会議等において、その研究成果が発表されている。サイドの違いにより、重きを置く点が異なっているためか、各サイドの交流は活発とはいえず、体系化はそれ程進んでいない。

典型的なテキストマイニングとしては、探索的なデータ解析とテキスト分類を挙げることができる。探索的なデータ解析は、テキストの中から未知の情報を発見したり、解答の得られていない問題に対する解答を発見したりする。一方、テキスト分類は、テキストに付随する何らかの分類基準に従って、テキストの集合をその分類基準ごとに分割する。テキスト分類自体は、新たな知見を発見するものではないものの、あるクラスに分類されたテキストの中に内在するパターンを発見したり、分類されたテキストの背後にある分布を発見したりすることができる。

3.2 テキストマイニング研究

テキストマイニングと一口にいても、テキストデータを対象として何らかのマイニングを実施していれば、テキストマイニングとみなすことができるため、多様な研究が存在する。以下においては、これらの研究のうちのいくつかの概要を紹介する。

ヘルシンキ工科大学のチームによる WEBSOM [17] は、テキストマイニングという用語が登場した初期の頃から開発が行われている。WEBSOM は、テキスト集合に含まれる単語が、対象とするテキストに含まれるかどうかでテキストを特徴付けることにより、SOM に基づいて各テキストを二次元上に配置する。このため、一度興味あるテキストを発見することができれば、その近傍を探索することにより、関連するテキストを発見することができる。また、WEBSOM では、最初に発見する興味あるテキストの探索を支援するために、与えられたキーワードに対して、フルテキストサーチを行う機能を提供している。

Bar-Iran 大学の R. Feldman [6] らによる FACT は、テキストデータに対して Apriori アルゴリズムを適用したツールである。FACT は、各テキストを特定のキーワードで特徴付けることにより、あるキーワード集合に対して、出現しやすいキーワードを識別する相関ルールを発見する。また、FACT では、支持度と確信度に基づいたルールの絞込みに加えて、キーワード

に定義する属性を利用した絞込みを行うことができる。さらには、FACT を発展させることにより、「名詞＋名詞」、「形容詞＋名詞」といった品詞の組み合わせを考慮したフレーズ間に存在する相関ルールを発見するツールの開発も行われている。

嶋津 [32] らによるコールセンター情報分析システムでは、係り受け関係にある 2 単語をパターンとすることにより、テキストの特徴付けを行う。このパターンをラベル付けされたテキスト集合に対して、Apriori アルゴリズムを適用することにより、パターン間の関係を示す相関ルールを発見する。本システムでは、意味のないルールの発見を抑えるために、事前確信度と事後確信度を導入し、この差の大きなルールを意味のあるルールとして抽出している。このとき、事前確信度とはルールの条件部を空集合とした場合の確信度であり、事後確信度とは通常の Apriori アルゴリズムにおける確信度である。

市村 [11] らによる日報分析システムは、対象とする日報に特徴的な表現を記述したキー概念辞書と表現の組み合わせにより生じる意味を記述した構造抽出ルールを利用することにより、入力されたテキスト集合を分類する。このとき、キー概念辞書は、概念クラス、キー概念、表層表現と呼ばれる 3 階層で構成されており、表層表現がテキスト中に実際に記述される表現、キー概念が同一の意味をなす表層表現をまとめた概念、概念クラスが関連するキー概念をまとめたものを表わしている。また、構造抽出ルールが、キー概念の組み合わせとその組み合わせにより発生する意味（クラス）を記述するルールを表わしている。本システムでは、この分類結果をクラスごとに集計することにより、テキストデータの全体的な傾向を分析することができる。また、店舗別、地域別等の条件を加えた分析を行うことができる。

櫻井 [30] らによるルール発見システムは、キー概念辞書を利用することにより、与えられたテキストに対する特徴付けを行う一方、利用者により与えられたテキストに対するクラスから、複数のキー概念を属性値、テキストクラスを分類クラスとする事例を生成する。このような事例を多数収集することにより、キー概念とテキストクラスとの間に内在するルールの発見を行う。また、学習されたルールを利用することにより、テキストクラスが割り当てられていない新たなテキストの分類を行う。本システムでは、ルールの発見に、ID3 をファジィ集合理論に基づいて拡張した IDF を利用しており、テキストに内在するあいまい性をファジィ集合を利用して表現している。

3.3 商用ツール

テキストマイニングは企業における動きも活発であり、多くのシステムが商用ツールとして提供されている。以下においては、そのうちのいくつかのツールを紹介する。

Concept Base はジャストシステムから提供されているツールであり、日本におけるテキストデータ分析ツールのはしりとなるツールである。単語のありなしでテキストを特徴付ける一方、テキストに対するクラスタリングや単語間の関連を視覚的にとらえる機能を備えている。また、形容詞と名詞との関係に着目した感性分析や、名詞と動詞との関係に着目した機能分析等の機能を提供している。

Vext Miner はクオリカ（旧 コマツソフト）から提供されているツールである。単語間の共起を利用したニューラルネットワークにより、類似テキストを近くに配置し、テキスト集合の関連を視覚的に把握する機能を提供している。SAS インスティテュートが提供するデータマイニングツールとの連携を図っており、データマイニングツールとシームレスにマイニング結果を受け渡しすることができる。

TRUE TELLER は野村総合研究所から提供されているツールである。50万語程度の辞書をサポートすることにより、高度な自然言語解析を実現している。テキスト集合に存在する単語のランキングや、単語の時間的な推移を追跡する時系列モニター、特徴的な単語を分析する機能等を備えている。

Knowledge Meister は東芝から提供されているナレッジマネジメントツールであり、その分析機能として、テキストマイニング機能が提供されている。タスクに対して記述されるキー概念辞書に基づいて、キー概念を含むテキストを抽出し、その件数を参照する機能を備えている。また、XML のデータベースとの連携を図っており、抽出したキー概念、テキスト集合のクラスタリング結果、テキストに付随する属性情報を組み合わせて、OLAP による分析を行う機能を備えている。

以上に紹介したツールの他にも、多数のツールが提供されており、機能の拡張も精力的に行われている。ただ、企業向けに開発されているツールが多いため、数百万円以上の導入費用が必要となるツールがほとんどであり、簡便に利用できるという状況にはなっていない。

3.4 研究のトレンド

テキストを特徴付ける方法としては、タスクに依存した辞書を利用しつつも、意味のある単語がテキストに含まれているかどうかを利用した研究が多くなされ

ている。構文解析や意味解析等のより高度な自然言語処理技術の研究も進んでおり、より高度な特徴に基づいて、何らかの知見を得るテキストマイニングが期待されている。

また、現在のテキストマイニングでは、ある時間断面で収集されたテキスト集合に対して分析を行う研究が多くなされている。営業日報の分析や流行の予測等においては、テキストの記述された日時を考量することも重要であり、時間経過を考慮したテキストマイニングが期待されている。

さらには、個人やタスクに依存した分析は、予め準備する辞書にその内容が反映される程度であり、誰にとっても簡便な分析が行える訳ではない。同一のテキスト集合であっても、利用者によって求める分析結果は異なっており、利用者やその目的にあった分析の実現に向けた支援が期待されている。

以上に述べた方向を含めて、種々の方向でテキストマイニング研究が発展していくものと考えられる。

4 Web 構造の分析

Web マイニングにおいては、ページ内のテキストデータだけではなく、それぞれのページがハイパーリンクでどのように結ばれているのかという Web 構造も、重要な分析対象となる。以下では、Web ページの間のリンク構造を解析するための代表的なアルゴリズムについて説明する。

4.1 PageRank アルゴリズム

PageRank アルゴリズム[24]は検索エンジン Google [9]のランキングの基本となっているアルゴリズムであり、リンク構造に基づいてページの重要性を決定しようというものである。

リンクに基づいてページの重要度を決定する場合、ページの被リンク数をカウントし、より多くのページからリンクされているページほど、重要度が高いとする方法が考えられる。しかしながら、この方法では、Yahoo!のような重要なページからのリンクも、個人のページからのリンクも同じ重要度で計算されてしまう。また、故意に大量のページを作成し、あるページへリンクを張ることで、ランキング値を操作するリンクスパムと呼ばれる行為に対して、脆弱であるという欠点がある。そこで、PageRank では、“重要なページから多くリンクされているページは重要である”という考えを採用している。ここで、ページの重要度は再帰的に定義されているが、これを以下のような反復計算で求める。

まず、ページ i にリンクしているページの集合を B_i

図1 PageRank のランダムウォークモデル

とし、ページ j がリンクしているページ数を N_j とする。このとき、ページ i の重要度 r_i を

$$r_i = \sum_{j \in B_i} \frac{r_j}{N_j} \quad (1)$$

を用いて、適当な初期値から始めて反復計算することで求める。これは、直感的には、リンク元 j が重要度の高いページであり、かつそのページが他のページにリンクしている数が少ないほど、 i の重要度に対する寄与が大きいことを意味している。

ここで、以下の m_w を要素とする行列 M を考える。

$$m_w = \begin{cases} \frac{1}{N_j} & \text{if } j \in B_i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

この行列を用いると、 r_i を要素とする PageRank ベクトルの $\vec{R}$ の計算式は、

$$\vec{R} = M \times \vec{R} \quad (3)$$

と書くことができる。すなわち、PageRank ベクトル $\vec{R}$ は行列 M の固有ベクトルであり、PageRank を計算するために、 $\vec{R}$ の適当な初期値から始めて行列 M を繰り返し掛けることは、べき乗法で行列の固有ベクトルを求めることに一致する。

PageRank の計算は、Web グラフ上でランダムウォークをするランダムサーファのモデルで捉えることができる。すなわち、 m_w をユーザがページ j からページ i へ遷移する確率と捉えれば、PageRank ベクトルは、推移確率行列 M のもとで、ユーザがページ i に滞在する平衡分布を与える。

しかしながら、 $\vec{R}$ が一意に収束することが保証されているのは、 M が既約（グラフが強連結で任意のページから任意のページへの経路が存在する）かつ非周期の

図2 Hub と Authority からなる Web コミュニティのモデル

場合である。実際の Web のリンク構造では一般に後者は成り立つが前者は成り立たない。そこで、PageRank アルゴリズムでは、dumping 項を導入し、

$$\vec{R} = (1 - \alpha) M \times \vec{R} + \alpha \vec{C} \quad (4)$$

としてページランクベクトルを計算している。ここで、 $\vec{C}$ の要素は全て $c_i = \frac{1}{N}$ である。（ N は全てのページの数。）

これは、図1のように、ランダムウォークをするユーザが、リンクをたどらない確率が α あり、その場合には、いずれかのページに等しい確率でジャンプするというモデルになっている。

これにより、任意のページの間パスが存在する（グラフが強連結である）ことになり、 $\vec{R}$ の収束は保証される。

dumping 項は当初は PageRank を計算する都合上導入されたが、特定のトピックやユーザに適応した PageRank を計算する際に、重要な役割を果たしている。

4.2 HITS アルゴリズム

HITS (Hyperlink-Induced Topic Search) アルゴリズム [14] は、図2のように、あるトピックに関して、権威的なページ (authority) と権威的なページへのリンクを多く含んだページ (hub) からなるページ集合を、Web 上のコミュニティとして抽出する方法である。各ページに対して、ユーザが投入したキーワードを含むページ、それらのページにリンクしてあるページ、それらのページからリンクされているページに基づいて、authority 値と hub 値を計算し、この値の大きなページを authority と hub として提示する。

このとき、authority 値と hub 値の計算は、行列の固

有ベクトルを求める問題に帰着することができるので、べき乗法や数値計算の分野で古くから研究されてきた固有ベクトルの計算法を利用することができる。しかしながら、Web におけるリンク解析は、膨大なページを対象とする必要があるため、より効率的に固有ベクトルを計算するアルゴリズムが望まれており、新たな研究課題として登場してきている[12]。一方、Web の構造解析を行う際に問題となるのが、抽出結果の安定性である。Web の構造は常に変化しているものの、少数のリンクの変化で、計算結果が大きく変化することは好ましくない。HITS は不安定であることが指摘されておりその改良アルゴリズムが提案されている [21]。また、リンク解析のみではなく、テキスト解析を利用して、アンカーテキスト周辺の情報を組み合わせた改良手法も提案されている [3]。

5 応用例

5.1 ポータルサイト作成の自動化

テキスト分類の応用例として、ポータルサイトの作成を自動化することが挙げられる。CORA [18] は、Web ページの中から計算機科学分野の論文を収集し、Naive Bayes Classifier を用いてトピックに従って階層的に分類を行っている。また、[8] では、単にページ内の単語だけでなく、リンク元のアンカー周辺の単語も収集して SVM を適用することで、精度のよい Web ページの分類と、カテゴリの命名の自動化が試みられている。さらに、WebKB [5] は単にページを分類するにとどまらず、オントロジーと関係学習の手法を用いて、プロジェクトのページとそのメンバーのページといった、Web ページの間のクラス関係を抽出し、知識ベースを構築することを試みている。

5.2 検索エンジンの高度化

Web 構造分析の最も成功した応用例は、検索エンジン Google であろう。テキスト解析や非リンク数を重視していた従来の検索エンジンでは、本文と関係のないキーワードをページ内に大量に入れることでランキングを上げようとするキーワードスパムや、故意に大量のリンクを張ることで、ランキングを上げようとするリンクスパムと呼ばれる行為に悩まされていたが、Google では Web のリンク構造を取り入れることで、この問題に対する頑健なランキングを提供した。

Google は汎用検索エンジンであり、PageRank アルゴリズムは、全ての Web ページに対して、1 つの PageRank ベクトルを計算し、グローバルな評価値を与えようというものである。これに対して、特定のトピックに特化した専門の検索エンジンや、個人の興味

に適応できる検索エンジンに対する要求も強い。例えば、同じ“architecture”というキーワードが投入された場合であっても、ユーザが建築のページに興味があるのか、あるいはコンピュータに興味があるのかによって、当然ページのランキングが変わってきて良いはずである。

そこで、PageRank アルゴリズムの dumping 項 $\vec{C}$ を適応パラメータとして用いることが提案されている[24]。Topic-Sensitive PageRank[10]は、それぞれのトピックごとに別々の PageRank を用意しておき、ユーザの想定しているトピックに応じて、ランキングを変化させようというものである。ここでは、Open Directory [22] の Art, Business, Computers といった16のトップカテゴリを利用し、カテゴリ k に属するページの集合を T_k とし、dumping 項を

$$c_{ki} = \begin{cases} \frac{1}{T_k} & \text{if } i \in T_k \\ 0 & \text{if } i \notin T_k \end{cases} \quad (5)$$

とすることで、PageRank を計算している。これは、ランダムウォークモデルでリンクをたどらずにジャンプを行う際、特定の集合に属するページにのみジャンプをすることに対応し、PageRank の計算においてトピックに応じた重み付けを行うことが可能になる。

また、検索エンジンの専門化を、テキスト解析の技術を用いて行うことも可能である。質問修正 (query modification) とは、ユーザが入力した質問に対して、適切な検索結果が返されるように、システムが自動的に修正を加える方式である。例えば、ユーザが料理のレシピを探す目的で“牛肉”というキーワードを投入した場合、システムが自動的に“材料”や“大さじ”といったキーワードを追加することで、検索精度を向上させることが可能である。決定木学習や SVM を用いて、このような質問の修正規則を抽出し、特定のトピックに適応した検索エンジンを構築することが行われている [7] [23]。

相関ルールを検索エンジンに用いた例としては、Mondou [13] が存在する。相関ルールを Web ページの集合から抽出する場合は、キーワードがアイテム、Web ページがバスケットに対応する。Mondou では、1 つのページ内でのキーワードの出現回数を考慮するために、重み付き相関ルールを導入している。

ユーザの入力したキーワードから、相関ルールを用いて関連するキーワードを導出し、ユーザに提示することで、検索の絞込みを支援している。

5.3 Web コミュニティ発見

HITS に基づく Web コミュニティ発見の研究は、IBM

のCleverプロジェクト[4]で行われた。HITS アルゴリズムは、ユーザが指定したキーワードに対応した Web コミュニティを、検索エンジンを用いて発見するというものであったが、Crawler を用いて Web ページを収集し、網羅的に Web コミュニティを発見することが行われている。Campfire[16]においては、Trawling と呼ばれるアルゴリズムを用いて Homeschooling, Hot and Spicy foods といった、100,000以上の Web コミュニティを発見することに成功している。また、[34]では、単に Web コミュニティを発見だけでなく、Web コミュニティの関連をも抽出し、ユーザが関連するコミュニティの間をナビゲーションすることを可能としている。例えば、コンピュータに関連するコミュニティとして、デバイスメーカ、ソフトウェアメーカ、コンピュータベンダといった複数のコミュニティが提示され、ユーザはそれらの間を巡回することが可能になっている。

6 おわりに

今回紹介した技術は、相関ルール、決定木、PageRank, HITS 等、どれをとっても基本的なアイデアは極めてシンプルなものである。しかしながら、Web は複雑かつ多様であり、実際の問題に適応する際には、様々なパラータ調整や改良が必要になる。また、Web 上の情報は、常に増加、多様化を続けているため、ある時点で有効であった技術が、その後も有効であり続けるという保証はない。

これは、情報の利用を支援するという立場からは、大きなチャレンジである。しかし、逆に言えば、多様で多量な情報があるからこそ、始めて実現可能になる手法もある。自然言語処理やテキストマイニングの立場から言えば、Web を巨大なコーパスとして活用することができるし、リンク解析は、一定数以上の Web ページがあるからこそ、有効であるといえる。

Web の上の情報が、単純な HTML テキストから、XML, マルチメディアデータ、セマンティック Web へと拡張していくにつれて、データマイニングの対象となる新たな問題、可能となる新たなサービスが今後も現れてくると考えている。

参 考 文 献

- [1] R. Agrawal, et al.: "Mining Association between Sets of Items in Massive Databases," *Proc. of SIGMOD Conf on Management of Data*, 207-216, 1993.
- [2] 赤穂 昭太郎: 「EM アルゴリズム-クラスタリングへの適用と最近の発展-」, 日本ファジィ学会誌, vol. 12, no. 5, 594-602, 2000.
- [3] S. Chakrabarti, et al.: "Automatic Resource List Compilation by Analyzing Hyperlink Structure and Associated Text," *Proc. of WWW7*, 1998
- [4] Clever Project: <http://www.almaden.ibm.com/cs/k53/clever.html>.
- [5] M. Craven, et al.: "Learning to Construct Knowledge Bases from the World Wide Web," *Artificial Intelligence*, vol. 118, no. 1/2, 69-113, 2000
- [6] R. Feldman and H. Hirsh: "Mining Associations in Text in the Presence of Background Knowledge," *Proc. of KDD'96*, 343-346, 1996.
- [7] E. Glover, et al.: "Improving Category Specific Web Search by Learning Query Modifications," *Proc. of SAINT2001*, 23-31, 2001.
- [8] E. Glover, et al.: "Using Web Structure for Classifying and Describing Web Pages," *Proc. of WWW2002*, 562-569, 2002.
- [9] Google: <http://www.google.com>.
- [10] T. H. Haveliwala: "Topic-sensitive Pagerank," *Proc. of WWW2002*, 517-526, 2002
- [11] Y. Ichimura, et al.: "Text Mining System for Analysis of a Salesperson's Daily Reports," *Proc. of PACLING 2001*, 127-135, 2001.
- [12] S. D. Kamvar, et al.: "Extrapolation Methods for Accelerating the Computation of Pagerank," *Proc. of WWW2003*, 261-270, 2003.
- [13] H. Kawano and T. Hasegawa: "Mondou: Interface with Text Data Mining for Web Search Engine," *Proc. of HICSS31*, 275-283, 1998
- [14] J. M. Kleinberg: "Authoritative Sources in a Hyperlinked Environment," *JACM*, vol. 46, no. 5, 604-632, 1999
- [15] T. Kohonen: "Self-Organizing Maps," Springer, 1995.
- [16] S. R. Kumar, et al.: "Extracting Large-scale Knowledge Bases from the Web," *Proc. of VLDB'99*, 639-650, 1999.
- [17] K. Lagus, et al.: "Websom for Textual Data Mining," *J. of Artificial Intelligence Review*, vol. 13, no. 5/6, 335-364, 1999
- [18] A. McCallum, et al.: "Automating the Construction of Internet Portals with Machine Learning," *Information Retrieval*, vol. 3, no. 2, 127-163, 2002.
- [19] T. Mitchell: "Machine Learning," McGraw-Hill, 1997
- [20] K. Nigam, et al.: "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, vol. 39, no. 2/3, 103-134, 2000.
- [21] S. Nomura, et al.: "Analysis and Improvement of Hits Algorithm for Detecting Web Communities," *Proc. of SAINT2002*, 132-140, 2002.
- [22] Open Directory Project: <http://dmoz.org>
- [23] S. Oyama, et al.: "Domain-specific Web Search with Keyword Spices," *IEEE TKDE*, 2003. (to appear)
- [24] L. Page, et al.: "The Page Rank Citation Ranking: Bringing Order to the Web," Technical report, Stanford Digital Library Technologies Project, 1998.

- [25] H. Pasula and S. J. Russell: "Approximate Inference for First-order Probabilistic Languages," *Proc. of IJCAI-01*, 741-748, 2001.
- [26] J. Platt: "Fast Training of Support Vector Machines using Sequential Minimal Optimization," MIT Press, 1998.
- [27] J. R. Quinlan: "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, 81-106, 1986.
- [28] J. R. Quinlan: "Learning Logical Definitions from Relations," *Machine Learning*, vol. 5, no. 3, 239-266, 1990.
- [29] 櫻井 茂明, 荒木 大: 「帰納学習によるファジィ決定木の生成」, 電学論 C, vol. 113, no. 7, 488-494, 1993.
- [30] 櫻井 茂明, 他: 「テキストマイニングシステム向けの構造抽出ルールの帰納学習」, 電学論 C, vol. 122, no. 6, 1009-1015, 2002.
- [31] A. D. Shapiro: "Structured Induction in Expert Systems," Addison-Wesley, 1987.
- [32] 嶋津 恵子, 他: 「相関ルール導出法によるコールセンター情報からの重要情報の発見」, 第17回人工知能学会全国大会, 1F4-03, 2003.
- [33] 鈴木 英之進: 「共通データからの仮説駆動型例外ルール発見」, 人工知能学会誌, vol. 15, no. 9, 782-789, 2000.
- [34] M. Toyoda and M. Kitsuregawa: "Creating a Web Community Chart for Navigating Related Communities," *Proc. of Hypertext 2001*, 103-112, 2001.
- [35] 馬野 元秀, 他: 「ID3に基づくファジィ・ルールの抽出とその推論方法」, 第9回ファジィシステムシンポジウム, 857-860, 1993.
- [36] V. N. Vapnik: "The Nature of Statistical Learning Theory," Springer, 1995.

(2003年8月20日 受付)

[問い合わせ先]

櫻井 茂明

〒212-8582 神奈川県川崎市幸区小向東芝町1

(株)東芝 研究開発センター 知識メディアラボラトリ

Tel : 044-549-2398

Fax : 044-520-1308

e-mail : shigeaki.sakurai@toshiba.co.jp

著者紹介

櫻井 茂明 [正員]

1991年東京理科大学大学院修士(数学)課程了。同年、(株)東芝入社。1998年から2年間新情報処理開発機構つくば研究センタ出向。入社以来、機械学習技術に関する研究開発に従事。現在、特にテキスト及び時系列データを対象とするマイニング技術を研究開発中。電気学会, 人工知能学会, 日本感性工学会各会員, 博士(工学)。

小山 聡 [非会員]

1994年京都大学工学部数理工学科卒業。1996年京都大学大学院工学研究科修士課程修了。1996年-1998年日本電信電話株式会社勤務。2002年京都大学大学院情報学研究科博士後期課程修了。博士(情報学)。現在、京都大学大学院情報学研究科社会情報学専攻助手。また、2003年よりStanford大学 Visiting Assistant Professor。機械学習, テータマイニング, 情報検索に興味をもつ。電子情報通信学会, 人工知能学会, 日本データベース学会, ACM, AAAI 各会員。