最適パターン発見にもとづく高速テキストマイニング

有村　博紀

1. 研究のねらい

現代の通信・計算機技術は、さまざまな事象に関する大量のデータの収集・蓄積を可能にした。しかし一方で、蓄積された大量のデータを、計算機の支援によって半自動的に解析し、有用な情報と知識をとりだす方法については、まだ十分には解明されていない。**データマイニング**は、データベースに蓄積された一見無意味にみえる大量のデータから、自明でない規則性やパターンを半自動的にとりだす方法についての科学的研究である。1990年代初頭から顕在化し、現在、理論と応用の両面で活発な研究が進んでいる。

一方、高速なネットワークと大容量記憶装置の発達を背景として、ウェブページに代表される**大規模テキストデータと半構造データ**の利用が急速に進みつつある。例えば、ウェブ検索エンジンで相互に結合されたウェブページの全体は、最大規模のテキストデータベースの例である。また最近では、ネットワーク上でのデータの流通・交換のための基盤技術として、XML などの半構造データ技術が盛んに研究されている。この膨大なデータ量の増加に対して、これらのテキストデータの内容に接近する方法としては、現在、情報検索などの限られた手段しかない。これらの**テキストデータからのデータマイニング**が緊急の課題となっている。

しかし、現在のデータマイニングの対象は、データが明示的で平坦な構造をもつ関係データベースが中心であり、これらのテキストデータベースは、

(1) 明示的な構造をもたない、

(2) 多様な電子化文書の、

(3) 膨大な量の集積

であることから、従来のデータマイニング手法をそのまま適用することができない。そのため、研究開始時点では、大規模テキストデータからのデータマイニングに関して、ほとんど研究がおこなわれていないのが現状であった。

そこで本研究では、大規模テキストデータを対象とした高速なデータマイニング・システムの研究を行なった。特に、データマイニングを、人間による大量のデータ解析を支援する効率的な半自動的ツールとしてとらえ、従来の情報検索システムを超えた新しい情報アクセスシステムの開発を目指した。とくに、計算量理論と計算学習理論との最新の成果を積極的に取り入れて、頑健かつ高速なアルゴリズムの開発を目標とした。

本研究でこだわった点は、第一に、実際の重要性にもかかわらず、あまり研究されてこなかったテキストデータベースからのデータマイニングを正面から研究対象としてとりあげ、新しい情報アクセスシステムの開発を目指すことである。第 2 に、データマイニングを自動的な知識獲得システムではなく、人間による大量のデータ解析を支援する効率的な半自動的ツールとしてとらえ、そのための技術と実現法を明らかにしようとする点である。第 3 に、計算効率に徹底的にこだわり、データ構造とアルゴリズム設計の技法を

駆使して、理論的に高速な(多項式時間の計算量をもつ)だけでなく、実際のデータに対して、きわめて高速なデータマイニング・アルゴリズムの開発を目指した点である。

2. 研究成果

本研究では、大規模テキストデータベースを対象としたテキストデータマイニングシステムの実現方式を明らかにすることを目指して研究を行なった。とくに、最適パターン発見に基づく高速なアルゴリズムに関して、理論・実装・応用の 3 つの観点から研究を行い、次のような研究成果を得た。

- 最適パターン発見に基づくパターン発見エンジンの開発。
- テキストマイニング・システムの大規模実装のための基盤技術
- テキストマイニングシステムのプロトタイプ構築と応用実験
- ウェブデータからの半自動的な情報抽出と、その計算複雑さの解析
- 半構造データを対象としたデータマイニング方式の開発

以下、各項目を詳細に述べる。

最適パターン発見に基づくパターン発見エンジンの開発

研究初年度の 1999 年から次年度の 2000 年度にかけて、テキストマイニングの要となる大規模テキストマイニングのための高速かつ頑健なパターン発見エンジンの開発に取り組んだ。

現在、テキストマイニングは盛んに研究されるようになっているが、研究を開始した時点では、テキストマイニングに関して、自然言語処理分野と情報検索分野で当時盛んであった(狭義の)テキストからの情報抽出(Information Extraction)または内容理解(Message Understanding)の研究がある程度で、大規模テキストデータを対象としたテキストマイニングの研究は、いくつかの先行研究を除いてあまり研究されていなかった。そこで、テキストデータマイニングとは何かを定式化するところから研究を開始した。

テキストデータに対する最も基本的な情報アクセス手段は、キーワード検索などの情報検索であると考えられる。ここで、情報検索とは、利用者が自分が興味をもっているテキストを特徴付けるようなキーワードを情報検索システムに与えて、テキストデータベース中でそのキーワードが出現するテキストの集まりを取り出すことをいう。そこで、テキストデータマイニングをこの**情報検索の逆操作**と定義することにした。つまり、テキストデータマイニングでは、与えられたテキストデータベースに対して、利用者は自分が興味をもつテキストの集まりを、事前の検索やいくつかのサンプルを示す等の何らかの手段で指定する。すると、情報検索とは逆に、テキストデータマイニングは、これらの興味深い文書の集まりをうまく特徴付けるキーワードやパターンを見つけ出すのである。

この立場に基づき、研究では、最適パターン発見に基づき、大量のテキストデータから、それらを特徴づけるパターンを発見するテキストマイニングツール AWAP(Algorithm for Word-Association Patterns)を開発した。 AWAP は最近の文字列アルゴリズム研究と計算学習理論を積極的に援用し、高速かつ頑健なパターン発見を実現している点に特色がある。

図１ ： 近接語相関パターンの出現例

語相関パターン： **(<attack on>, <oil platform>; 4)**

- Monday 's **attack on** two Iranian **oil platform** by American forces in the ...
- ...the action would involve an **attack on** an **oil platform**
- The **attack on** the **oil platform** was the latest example of a U.S....
- ...**attack on** an Iranian **oil platform** in the Gulf on Monday appeared to ...
- A top Iranian military official said America 's **attack on** an Iranian **oil platform** on Monday
- had involved the United States in full-scale war ...

基本的に、開発したテキストマイニング・システム AWAP は入力テキスト上で、与えられた分類に関する統計的指標を最適化するパターンを発見するアルゴリズムである。直感的には、経済関係の新聞記事の集合から、海運関係の記事を特徴付けるパターンを見つけようとしている場合、AWAP は海運関係の記事により多く出現し、海運関係以外の記事にはあまり出現しないパターンを高速に見つける。ここで、分類に関する統計的指標とは、パターンがテキスト集合 POS をテキスト集合 NEG からうまく分離する度合いを示す関数であり、情報エントロピーや分類精度、Gini 指標、χ 二乗指標等を用いる。

この枠組みは、**最適パターン発見**(optimized pattern discov-ery)と呼ばれ、比較的単純なパターンのクラスから、入力データ上で与えられた統計的尺度を最小化または最大化するパターンを見つけることをいう。最適パターン発見は、1970 年代の統計的決定理論にその源をもち、1990 年代に入って、計算学習理論やデータマイニング分野で再発見された。

この最適パターン発見の枠組みを採用し、統計的指標を最適化するパターンを見つけることで、ノイズや不完全データに対する頑健性が高まる。その一方で、パターン発見問題の計算量が大きくなるという問題が生じる。そこで、AWAP は、高速にパターンを計算するために、テキストのもつ組合せ構造をうまく利用することで、大量のテキストに対して、実際に線形に近い時間で動くように設計されている。

AWAP は、パターンの族として、**近接語相関パターン**(proximity word-association patterns)と呼ばれる単純な文字列パターンの族を用いる。近接語相関パターンとは、テキストの部分語のリストと正整数からなる (″attack on″、 ″oil platform″、 4) のような表現である。これは、テキスト中で文字列 attack on の後に、oil platform が、出現の先頭間が 4 単語以下の距離(近接度という)しかはなれずに、与えられた順序で出現するという制約を表している。一般には、近接語相関パターンは、非負定数 d、 k に対して、d 個の任意長の文字列(この場合は単語列) Ai と近接度 k の組からなるパターン ($A1$、 …、 Ad; k) である。図 1 に、実際にテキストマイニング実験でみつかった近接語相関パターンとそのテキスト中の出現の例を示す。

近接語相関パターンに対する最適パターン発見問題は、語数 d と近接度 k が定数で制限されるときは、全ての可能なパターンを枚挙し、それぞれに対してテキストを走査し、出現数を調べる単純な生成枚挙法を用いて $O(n^{2k+1})$ 時間で解ける。これは、理論的には多項式時間アルゴリズムであるが、多項式の次数が大きく、現実には大規模テキストデータには適用できない。そこで、この**近接語相関パターンのクラスに対して、高速に最適解を見つけるアルゴリズムを開発**した。

このアルゴリズムは、接尾辞木(suffix tree)と呼ばれるフルテキスト索引構造を用いて、テキスト中のすべての部分文字列をその文字列の内容で組織化し、文字列の内容で定まる順位空間と文字列の出現位置

で定まる位置空間をいったり来たりしながら、テキスト中に実際に出現するパターンだけについて、そのパターンの形とその出現位置全体の集合を同時に計算していく。また、計算幾何学的な観点からは、このアルゴリズムは、テキスト上のパターンの発見を順位空間上の超直方体からなる幾何学的パターンの発見問題に帰着しているともみなせる。この手法を用いることで、このアルゴリズムは、サイズ N のランダムテキスト(現実の英文テキストや遺伝子配列テキストはランダムに近い性質をもつ)から、**最適な k 近接 d 語相関パターンを線形に近い時間で計算する**ものである。これにより、十分な主記憶が利用できる場合には、このアルゴリズムを用いて高速な最適パターン発見が可能になった。

高速な最適パターン発見アルゴリズムは、幾何学的パターンの発見に対しては従来知られていたが、テキストパターンに対する高速な最適化アルゴリズムはおそらくはじめてのものであると考えられる。

テキストマイニング・システムの大規模実装のための基盤技術

接尾辞配列を用いた実用的な高速パターン発見アルゴリズム。最初に開発したアルゴリズムは、ランダムテキストに対してほぼ線形に近い計算時間をもち、理論的には高速なアルゴリズムである。しかし、予備的な実装と実験によって、実際のデータ上では計算時間と使用領域量の両面で非効率的であり、現実の大規模データには適用不能であることがわかった。この非効率性の主な原因は、パターン発見アルゴリズムの主要なテキスト索引構造として用いている接尾辞木に問題があることがわかった。

接尾辞木は、主記憶上に構築される一種の木構造のデータ構造である。そのため、(1)主記憶上の木構造であり、メモリアクセスパターンが秩序的でないためキャッシュが効きにくい、(2)準線形時間の動作を保証するための索引構造の動的再構成を高速に行なうことが難しい、(3)木構造は領域量が大きい、(4)アルゴリズムの構造が複雑でコードチューニングが難しいことなどの問題を生じていた。

アルゴリズムの鍵は、テキスト中の部分語の効率よい枚挙とその出現位置の高速な計算である。従来、これらの操作の実現は、接尾辞木を用いるしかないと考えられていた。本研究では、接尾辞木の代わりに、接尾辞配列(suffix array)と高さ配列(height array)という1 次元整数配列を組み合わせて、**より実装しやすく、記憶効率が良いマイニング向けテキスト索引技法を開発**した。このために必要な技術として、接尾辞配列の一方向走査を用いた接尾辞木の巡回の模倣や、索引構造の動的な再構成法、当時知られていなかった接尾辞配列からの高さ配列の線形時間構築法等の技術からなっている。

これらの改良により、AWAP の適用可能な規模と計算時間を著しく改善することができた。例えば、MB の主記憶を搭載した計算機上で、従来技術では、50KB 程度のテキストに対して数十分以上必要としたのが、10 数 MB のテキストに対して数分以下で計算可能になった。これらの技術は、AWAP の実装だけでなく、テキストマイニングに必要な基本的な演算を実装するため基本的な技術として、さまざまなテキストマイニング問題に利用可能な技術である。

図2 : AWAPの出力例

(a)AWAP (エントロピー最小化)

順位	パターン
1	<shipping>
2	<ships>
3	<gulf>
4	<vessels>
5	<the gulf>
6	<port ship>
7	<ship>
8	<kuwaiti>
9	<iranian>
10	<iran>
11	<in the gulf>
12	<tankers>
13	<cargo>
14	<vessel>
15	<warships>
16	<strike>
17	<attack>
18	<tanker>
19	<flag>
20	<ports>

順位	パターン
81	<strait>
82	<strait of hormuz>
83	<chinese made>
84	<kuwaiti oil>
85	<caspar>
86	<iran and>
87	<oil platforms>
88	<gulf to>
89	<shipping sources said>
90	<attacked>
91	<the attack>
92	<gulf the>
93	<ferry>
94	<in rio de janeiro>
95	<lloyds shipping intelligence>
96	<the strait of hormuz>
97	<weinberger said>
98	<the waterway>
99	<the strait>
100	<bulk>

b)従来手法(頻度最大化)

順位	パターン
1	<reuter>
2	<the>
3	<to>
4	<said>
5	<of>
6	<and>
7	<in>
8	<a>
9	<s>
10	<on>
11	<for>
12	<at>
13	<by>
14	<said the>
15	<in the>
16	<with>
17	<from>
18	<of the>
19	<was>
20	<but>

外部記憶指向の高速パターン発見アルゴリズム。昨年度開発した Split-Merge アルゴリズムは、大量の主記憶を必要とするために、外部記憶におかれた数百メガバイトを超えるような大量のテキストデータに対しては、この方式をそのまま拡張することができない。また、ウェブデータからのパターン発見においては、順序制約がない近接語相関パターンの方が適している場合も多いしかし、Split-Merge アルゴリズムでは、パターンの探索において語の順序情報が本質的であり、順序なしの近接語相関パターンに拡張することが難しい。これらの問題に対して、関係データベース向けのデータマイニングアルゴリズムとして広く用いられている Apriori アルゴリズムが採用しているディスク走査のアイディアに基づいて、外部記憶指向の高速な発見アルゴリズムを開発した。この結果は、PAKDD2000 国際会議で発表し、同会議の優秀論文賞を受賞した。

テキストマイニングシステムのプロトタイプ構築と応用実験

開発したアルゴリズムとデータ構造を元に、テキストマイニングのプロトタイプシステムを構築し、動的文書ブラウジングとウェブマイニングに適用する実験をおこなった。

従来のテキストマイニング方式では、頻度の高いフレーズを抽出する単純な手法を採用している。ものが多い。しかしこの方法では、対象データに特徴的なフレーズが、自明な高頻度語に隠蔽されてしまい、発見されないことが多い。また、伝統的な高頻度語の除去も有効でない。そこで、利用者が興味をもっているテキストを正例とし、それ以外のテキスト全体を負例として、最適パターン発見アルゴリズムを適用することで、特徴的なパターンを発見することをこころみた。

図 2 の(a)に、海外貿易に関する英文新聞記事データから、エントロピー最小化を用いて、AWAP が見つけた部分語パターン(語数 1)を示す。順位 1～20 位には、主題に強く関連するキーワードが、81～100 位にはある種の要約とみなせる低頻度の長いパターンが見つかっている。図2の(b)には、従来の頻度最大化で見つかる自明なキーワードを示す。

図3：ウェブマイニング実験

(a) HONDA vs. SOFTBANK

Rank	Pattern	Rank	Pattern
1	<honda>	11	<miles>
2	<prelude>	12	<bike>
3	<i>	13	<motorcycle>
4	<car>	14	<racing>
5	<parts>	15	<black>
6	<engine>	16	<si>
7	<99>	17	<me>
8	<rear>	18	<tires>
9	<vtec>	19	<fuel>
10	<exhaust>	20	<my>

(b) HONDA vs. TOYOTA

Rank	Pattern	Rank	Pattern
1	<honda>	11	<honda prelude>
2	<prelude>	12	<98>
3	<vtec>	13	<valkyrie>
4	<si>	14	<99 time>
5	<bike>	15	<honda s>
6	<99>	16	<rims>
7	<motorcycle>	17	<looking>
8	<the honda>	18	<looking for>
9	<prelude si>	19	<scooters>
10	<civic>	20	<black>

図3に、ウェブ検索エンジンとウェブロボットを用いて収集したウェブページ集合を対象にしたテキストマイニング実験の結果を示す。ここでは、情報エントロピーを用いて、自動車会社 HONDA 関連のページを特徴付けるパターンを発見するのが目標である。データは、正例と負例ともに、タグを除いて 5MB 前後である。図 2 で、自動車に関連しないページと比較した左側(a) では、一般的な自動車用語が抽出されている。同じ自動車会社 TOYOTA と比較した右側(b) では、HONDA が生産している具体的な車種名が抽出されていることがわかる。非常に多い負例に対してもうまく働く。最適パターン発見が、語彙や内容に関する事前知識なしに、適切なキーワードを見つけていることに注意されたい。

なお、本研究に関する論文(人工知能学会誌 Vol。15、 No。4、 2000 年)は、人工知能学会 2000 年度論文賞を受賞した。

ウェブデータからの半自動的な情報抽出と計算複雑さ

研究目標の微調整。 初年度からの 1 年半の研究成果により、当初挙げた項目については、テキストマイニングの基本的な枠組みを確立できる見通しがついた。一方、テキストマイニングに関するこの 1 年半の研究によって、ウェブテキストや XML データなどの半構造データからのデータマイニングが、緊急かつ重要な課題として顕在化しつつあることがわかってきた。具体的には、

- ウェブテキストや XML データなどのネットワーク上に分散した巨大かつ多様なテキストデータからの知識獲得が、これからのテキストマイニングの焦点となる。
- 構造をもつ半構造データを効率よく扱うための基本的な技術は、現在、未成熟である。

そこで、研究後半に入る平成 12 年下半期からの 1 年半で、とくに半構造データからのデータマイニングに焦点を絞り、「半構造テキスト＝テキスト＋構造」ととらえて、対象をテキストから木構造をもつ複合データである半構造テキスト(半構造データ)に拡大して、当初の計画をつづけて実施していくことにした。

ウェブデータからの情報抽出。 半構造データからの情報獲得では、従来のデータマイニングとテキストマイニングの枠組みを超えた多様な問題が生じる。一般に、木(tree)は 2 次元の対象であり、1 次元の対象であるテキストに対するアルゴリズムが直接木構造データに効率よく拡張できないなど、固有の問題が生じる場合も多い。そこで、手始めとして、半構造データからのパターン発見問題のケーススタディとして、ウェブデータからの情報抽出アルゴリズムの開発にとりくんだ。

ウェブデータからの情報抽出の研究では、半構造データの形式的なモデルとして、ラベルつきの順序木を採用し、最小汎化と呼ばれる機械学習手法を、与えられたウェブページとそこからの切り出し例から、ウェブページからの切り出しを行なう HTML ラッパーを自動的に構築するアルゴリズムを考案した(ICGI'02)。さらに、これを、一般的な属性やタグをもつ HTML データに拡張し、実際のウェブページを用いた実験では、先行研究である Kushmerick のラッパー構築アルゴリズムより安定した切り出しを行なうことを確認できた(DS'00)。

さらに、上記の木構造ラッパー帰納問題や、構造付き入れ子テキストパターンの学習、半構造パターンの対話的学習など、さまざまな構造パターン発見問題に関して、パターン発見の本質的な計算複雑さの解析を行ない、パターン発見問題の計算複雑さを同定した(ICGI'00、 ALT'00、 TCS 2000)。これは、本質的に困難な問題に対してやみくもに経験的アルゴリズムを与えることを避けることを意図したものである。先の半構造データマイニングの目標とするパターンの族の同定に役立った。

半構造データマイニング・アルゴリズムの開発

最終年度である平成 13 年上半期から 14 年上半期にかけて、「特徴的なパターンの発見」に関して、半構造データを最も基本的なラベル付き順序木のクラスとしてモデル化し、データ中の頻出共通部分構造に対する高速な発見アルゴリズムを開発した。

一般に、木に関するパターン発見問題は、高い計算量をもつことが多い。そこで、最右枝拡張法という効率よい発見手法を与え、これを複数の最適化手法と組み合わせて、半構造データに対する高速なマイニングアルゴリズムを与えた(SDM'02)。さらに、開発したアルゴリズムを、最適パターン発見の枠組みに適用し、目的属性で 2 値ラベル付けされた半構造データから、目的属性に関する情報エントロピー関数値を最小化する部分構造を高速に発見するアルゴリズムを与えた(PKDD'02)。これにより、当初の目的である最適化パターン発見に基づいた半構造データマイニングについて、基本方式を与えることができた。本研究に関する国内ワークショップ(DEWS'02) での発表は、2002 年 5 月に同ワークショップ優秀論文賞を受賞した。

3. 今後の展開

さきがけ研究での最大の目標であった、高速かつ頑健なテキストマイニングシステムについては、そのための基本的な技術の開発に成功し、プロトタイプシステムによる実験においても基本的な応用可能性を示すことができた。今後は、この技術を、ネットワーク上にあふれる膨大な量の情報から、自分が必要とする情報を得るための道具とするための研究も進めたいと考えている。そのためには、高速なアルゴリズムの開発だけでなく、開発した技術を実際の人間の活動にどのように埋め込んでいくべきかといったソフトウェアのデザインの問題や人的因子の問題も考えていくべきかもしれない。今後は、他分野・領域外の専門家との協同研究も進めながら、こだわっていきたい。

半構造データマイニングについては、現在、半構造データに対して効率の良いデータマイニング・アルゴリズムの設計が可能であることを、きわめて単純な問題に対して示すことができたという段階である。半構造データマイニング研究をどのように展開させていくべきかさらに探求をする必要がある。現在、ネットワークの侵入検出や、XML データストリームからの情報獲得を目標として、上記の半構造パターン発見ア

ルゴリズムをもとに、半構造データストリームからの効率よいデータマイニング手法を開発したところである(ICDM'02、 to appear)。現在、データマイニング分野では半構造データマイニングに関して、さまざまな技術提案や応用可能性が出されており、さらに継続して研究を進めたい。

4. 成果リスト

論　文

1. T. Asai, H. Arimura, K. Abe, S. Kawasoe, and S. Arikawa, Online Algorithms for Mining Semi-structured Data Stream, Proc. IEEE International Conference on Data Mining (ICDM'02), IEEE Computer Sciety Press, December 2002. (To appear)
2. K. Abe, S. Kawasoe, T. Asai, H. Arimura, S. Arikawa, Optimized Substructure Discovery for Semi-structured DataProc. 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2002), LNAI 2431, Springer-Verlag, 1-14, August 2002.
3. T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, S. Arikawa, Efficient Substructure Discovery from Large Semi-structured Data, Proc. Second SIAM International Conference on Data Mining (SDM'02), 158-174, SIAM, 2002.
4. H. Sakamoto, K. Hirata, and H. Arimura, Learning Elementary Formal Systems with Queries, Theoretical Computer Science, 2002. (accepted)
5. 村上義継、坂本比呂志、有村博紀、有川節夫(九大)、HTML からのテキストの自動切りだしアルゴリズムと実装、情報処理学会論文誌 ： 数理モデル化と応用、Vol. 42, No. SIG 14 (TOM 5), 39-49, Dec 2001.
6. 安積裕樹、川副真治、安部潤一郎、有村博紀、有川節夫(九大)、分散記憶型並列計算機における大規模接尾辞配列の構築法、情報処理学会論文誌 ： 数理モデル化と応用、Vol. 42, No. SIG 14 (TOM 5), 14-24, Dec 2001.
7. H. Arimura, H. Sakamoto, S. Arikawa, Efficient Learning of Semi-structured Data from Queries, Proc. the 12th International Conference on Algorithmic Learning Theory (ALT'01), LNAI 2225, 315-331, Springer-Verlag, 2001.
8. K. Taniguchi, H. Sakamoto, H. Arimura, S. Shimozono and S. Arikawa, Mining Semi-Structured Data by Path Expressions, Proc. the 4th International Conference on Discovery Science, LNAI 2226, 378-388, Springer-Verlag, 2001.
9. A. Yamamoto, K. Ito, A. Ishino, H. Arimura, Proc. the 11th International Conference on Inductive Logic Programming (ILP'01), LNAI 2157, Springer-Verlag, 2001
10. T. Kasai, G. Lee, H. Arimura, S. Arikawa, K. Park, Linear-time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications, Proc. the 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01), LNCS 2089, 181-192, Springer-Verlag, 2001.
11. H. Arimura, H. Asaka, H. Sakamoto, S. Arikawa, Efficient Discovery of Proximity Patterns with Suffix Arrays (Extended Abstract), Proc. the 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01), Short talk, LNCS 2089, 152-156, Springer-Verlag, 2001.
12. H. Sakamoto, H. Arimura, and S. Arikawa, Extracting Partial Structures from HTML Documents,

Proc. the 14th Florida Artificial Intelligence Research Symposium (FLAIRS'2001), Florida, AAAI, 264-268, May, 2001.

13. H. Arimura and S. Jain (eds.), Proc. the 11th International Workshop on Algorithmic Learning Theory (ALT'00), LNAI 1968, Springer-Verlag, Sydney, Dec. 2000.
14. H. Arimura, J. Abe, R. Fujino, H. Sakamoto, S. Shimozono, S. Arikawa, Text Data Mining: Discovery of Important Keywords in the Cyberspace , Proc. Kyoto International Conference on Digital Libraries 2000, Kyoto University, British Library and National Science Foundation (U.S.A.), 121-126, 2000.
15. H. Sakamoto, H. Arimura, S. Arikawa, Identification of Tree Translation Rules from Examples, Proc. the 5th International Colloquium on Grammatical Inference (ICGI 2000), LNAI 1891, Springer-Verlag, 241-255, Sep. 2000.
16. 安部 潤一郎、藤野 亮一、下薗 真一、有村 博紀、有川 節夫、テキストデータからの高速データマイニング人工知能学会誌、Vol.15, No.4, 2000 年 7 月
17. H. Arimura, H. Sakamoto, and S. Arikawa, Learning Term Rewriting Systems from Entailment, 10th International Conference on Inductive Logic Programming (ILP2000) Work-in-Progress paper session, July 2000.
18. H. Arimura, Text Data Mining with Optimized Pattern Discovery, Proc. the 17th Machine Intelligence - Life Long Learning and Discovery in Procedural and Declarative Knowledge, K. Furukawa, S. Muggleton, D. Michie, and L. DeRaedt (eds.), Bury St. Edmunds, UK, 19 - 21July 2000.
19. R. Fujino, H. Arimura, S. Arikawa, Discovering Unordered and Ordered Phrase Association Patterns for Text Mining, Proc. 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2000), LNAI 1805, 281-293, Springer-Verlag, Nara, Apr. 2000.
20. S. Shimozono, H. Arimura, and S. Arikawa, Efficient Discovery of Optimal Word-Association Patterns in Large Text Databases, New Generation Computing, 18, 49 - 60, 2000.
21. A. Yamamoto and H. Arimura, Inductive Logic Programming : From Logic of Discovery to Machine Learning, Special Issue on Surveys on Discovery Science, (Eds.) S. Miyano, IEICE Transaction on Infomation and System, E83-D (1), 10-18, 2000.

解説記事

1. 有村 博紀、坂本比呂志、データマイニングにおける最適パターン発見、応用数理、応用数理学会、2002。(予定)
2. 池田 大輔・坂本 比呂志・有村 博紀、ウェブデータマイニング、システム/制御/情報「データマイニング特集号」、システム制御情報学会、第 46 巻第 4 号、Apr. 2002.
3. 坂本比呂志、有村博紀、Web マイニング、特集「テキストマイニング」、人工知能学会誌, Vol. 16, No. 2, 2001 年3月.
4. 那須川哲哉、河野浩之,、有村博紀、テキストマイニング基盤技術、特集「テキストマイニング」、人工知能学会誌、Vol. 16, No. 2, 2001 年3月.

受　賞

1. 電子情報通信学会 DE 研究、DEWS2002 優秀論文賞、2002 年5月受賞。
2. 人工知能学会 2000 年度論文賞、2001 年 5 月受賞。
3. PAKDD2000 Paper with Merit Award, 2000 年 4 月受賞。
4. 人工知能学会 1999 年度全国大会優秀論文賞、1999 年 12 月受賞。

招待・依頼講演

1. H. Arimura, Efficient Text Mining with Optimized Pattern Discovery (invited talk), Proc. the 13th Annual Symposium on Combinatorial Pattern Matching (CPM'02), LNCS 2373,17-19, Springer-Verlag, Fukuoka, July 2001.
2. 坂本比呂志、村上義継,安部潤一郎、有村博紀、有川節夫、ウェブからの情報抽出と最適パターン発見、特別セッション「データ・テキストマイニングにおける統計的モデリングの実際」、第 4 回情報論的学習理論ワークショップ(IBIS2001),117-122, 2001.
3. 有村博紀、最適パターン発見にもとづくデータマイニング、統計数理とデータマイニング、統計数理研究所共同研究リポート、142,13-24, 統数研、2001。
4. 有村博紀、データマイニング－ウェブデータからの知識発見を目指して－、電子情報通信学会 IT 研究会、2000。
5. チュートリアル企画、「発見科学とデータマイニングの最前線 ： 金融・経済・ゲノムからウェブまで」、2001 年電子情報通信学会ソサイエティ大会、2001 年 9 月。
6. 有村博紀、テキストマイニング ： ウェブデータからの知識発見を目指して、第25回情報化学討論会概要集、J13、日本化学会情報化学部会、2002。(予定)