

感情表現の抽出手法に関する提案

中山 記男^{†‡} 江口 浩二^{†‡} 神門 典子^{†‡}

[†] 総合研究大学院大学 情報学専攻 〒101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所内

[‡] 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: [†] norio@grad.nii.ac.jp, [‡] {eguchi, kando}@nii.ac.jp

あらまし Blog 形式の Web サイトが増加したことや、Amazon.com のようなオンラインショッピング利用者が増加したことによって、Web 上には製品や事象への評価情報が多く見られるようになった。それにより、評価情報を用いた意思決定を支援する手法への関心が高まっている。本稿ではその評価情報の中でも「うれしかった」「悲しかった」などの感情表現に着目し、抽出方法のモデル化と抽出システムの提案を行う。提案手法では、機械学習の手法であるブートストラップとシーケンシャルパターンマイニングの手法である Prefixspan を組み合わせることによって半自動的に多様な長さの文字列で表現される感情表現を抽出し、辞書構築を行った。その結果、収集された感情表現とその評価を報告する。

キーワード 情報抽出, 感情抽出, 評判情報, 意見抽出, 感情表現, テキストマイニング, ブートストラップ

A Proposal for Extraction of Emotional Expression

Norio NAKAYAMA^{†‡} Koji EGUCHI^{†‡} and Noriko KANDO^{†‡}

[†] Department of Informatics, The Graduate University for Advanced Studies, Tokyo, 101-8430 Japan

[‡] National Institute of Informatics, Tokyo, 101-8430 Japan

E-mail: [†] norio@grad.nii.ac.jp, [‡] {eguchi, kando}@nii.ac.jp

Abstract Many online reputation sites using evaluative expressions such as blog web site or online shopping site like Amazon.com has been emerged on the Web. The interest for decision support mechanism using online reputation has been increased. We focused our attention on emotional expressions like “happy” or “sad” within many types of evaluative expressions. In this paper, we propose an emotion extraction method and present a prototype system based on this method. In this proposal, we explored the semiautomatic process to produce emotion lexicon collecting flexible length emotional expressions by combining bootstrapping method and Prefixspan which is the technique for a sequential pattern mining. Finally, we report evaluative expressions in the lexicon collected with this process.

Keyword Information Extraction, Emotion Extraction, Reputation, Opinion Extraction, Emotional Expression, Text Mining, Bootstrap

1. はじめに

本研究では Web サイトのテキストを対象として、そこに含まれる感情表現の抽出および辞書構築を行った。

近年注目されているものとして、Web 上のテキストに記述された評価情報がある。これには、例えば個人が自身の Web サイトにおいて事象や製品に対する評価を述べたり、または Amazon.com をはじめとする、Web 上の仮想店舗でユーザが製品に対し評価を述べることのできるケースがある。個人の Web サイトにおける評価情報は、Blog という専用のツールを用いることで、事象や製品にコメントしていくサイト形態の増加によって爆発的に増加し、また Web 上の仮想店舗にあるような評価情報も、店舗の利用者が増加することによって相対的に増加しつつある。このような評価情報が

増加したことにより、ユーザはそれを参考にして購入を、開発者はそれを参考にした開発をするといった利用がされはじめている。これに伴い、Web 上に存在する評価情報を積極的に活用することに対する要求が高まりつつある。しかしながら評価情報が記述されたテキストの数は膨大であり、人手での処理には時間的に限りがあるため、なんらかの手法を用いた自動的な処理が必要である。ただし、Web ページの表現形態や評価情報の表現形式が不定形であるため、予め定義されたルールに基づいて処理するのは現実的ではない。

以上のような問題に対する処理手法のひとつとして、情報抽出がある。情報抽出とは、テキストから特定のイベントや事柄に関する情報を自動的に抽出することである。人名や事柄、期日などの抽出する項目を

あらかじめ定め、そして機械学習やパターンマッチングによりその項目を抽出する技術が典型的な情報抽出である。

本研究では、この情報抽出の技術を応用し、評価情報の中でも「悲しかった」「うれしかった」「あたまにきた」等の感情表現に着目し、自動的に抽出する方式について検討を行った。感情表現を抽出・分析することにより、ユーザがその対象とする製品・事象についてどのような感情を抱いたのかを知ることができる。この技術のひとつの利用用途として、評価情報を持つ Web 上の仮想店舗に適用し「悲しくなる映画」「元気の出る音楽」「怒りを覚える本」のような検索を可能とすることがある。特に映像や音楽・図書作品や事象などに対する評価は、そのほとんどが主観的なものであり、その評価を含む感情を同定することが評価からの作品検索を手助けするとも考えられる。

従来の検索技術では、検索対象のテキストや索引付けされたジャンルなどの属性によって検索が行われることが多かった。そのため作品のタイトルや紹介文に含まれる語やジャンルから検索することは可能でも、「うれしかった」「悲しくなった」などの評価から作品を検索する方法は限られている。例えば画像や絵画を感性語で検索する手法では、事前に索引付けた画像に関する属性（情報）を用いて検索が行われることが多く、評価情報を用いての検索とは手がかりが異なる。

他の利用用途として、感情によるメールのフィルタリングも考えられる。緊急性の高い苦情メールの優先表示や、自分が発信するメールに不要な感情が含まれていないかどうかをチェックすることができる。

近年、同じく情報抽出の分野で評判情報検索[1]と呼ばれる研究がある。評判情報検索とは、テキストからその事柄に関する評判を検索する技術である。この検索された評判から意見を抽出し、事柄自身の評価に用いる。この際対象となるものは、ある単一のモノの評判である場合もあれば、特定の話題分野といった大きな枠組みの中での評判が対象となる場合もある。この研究以外にも、評価表現抽出や、評価表現の肯定・否定分類など、Web 上に存在する評価からなんらかの知見を得ようとする関心は高まってきている。これら関連研究と本研究との関係については、次節にて述べる。

本稿では情報抽出の応用として、テキストから感情表現を抽出するためのルールを半自動的に作成し、かつそれを用いて感情表現を自動的に抽出し、辞書を構築する方法について焦点を当てている。2 章では関連研究、3 章では抽出のシステム、4 章では抽出され構築された辞書に関する評価、そして 5 章ではまとめとともに今後の課題を示す。

2. 関連研究

本節では本研究の関連研究について述べる。言及する関連研究を大きく 3 つに分類した。本研究が扱う評価情報中の感情と、関連研究で扱われるその他の評価情報の関係を表 1 に示す。

表 1 評価情報の分類

	情報の性質	情報の種類	実例
主観的 情報	主観に 基づく 評価	感情	うれしかった
		評判	ブレーキ性能 が良い
		意見	ここを直して ほしい
客観的 情報	事実 に基づく 評価	事実報告	最高時速 300 km/h
		比較	A 車より B 車の ほうが高い

2.1. 意見・評価の収集に関する研究

Riloff ら[10]は、新聞記事を対象とし、ブートストラップを用いて主観的な意見を取り出すことに着目している。館野[7]は企業に寄せられる「お客様の声」を対象とし、文の構造から抽出規則を作成することで、不満足・否定を持つような緊急性の高い表現を抽出する方法について着目している。これらの研究は、パターン学習や抽出ルールを用いて、対象とするテキストから目的とする表現を抽出している。これらはそれぞれ主観的な名詞、不満足などを示す特定の表現を抽出することを目的としており、本研究のような感情表現のみの抽出を主眼としたものとは、抽出対象が異なる。

2.2. 意見・評価の分類に関する研究

Dave ら[2], Pang ら[3], Turney ら[8], Wilson ら[9], 立石ら[1], Kobayashi ら[4], Nasukawa ら[6]は、特定の製品や作品に対する評価表現を収集し、かつそれが肯定・否定・その他のどれかに分類することに着目している。特に Wilson らは意見の強さにも着目しており、立石らは特定の製品ごとの分類項目を用い、その製品に見合った評価表現での比較を行っている。これら研究はあらかじめ決められたカテゴリー（主に肯定と否

定)に評価を落としこむことを主眼としており、その評価の内容や感情にまでは触れられていない。肯定と否定で分類することが有効な分野には適しているが、映像や音楽、図書などのように評価が多様で肯定・否定の2値では扱えないような分野では、例えば「泣けた」「元気が出た」「あたまにきた」など、それ以外の評価を扱うことも必要である。

2.3. 感情の付与に関する研究

Liu ら[5]は、常識知を用いてメールの本文テキスト中に含まれる感情を抽出し、あらかじめ決められた感情カテゴリーへ分類することでテキストの感情を理解することに着目している。田中ら[12]は、結合価パターンを用い、日記中の文の情緒推定を行うことで、辞書へ情緒生起情報付与することに着目している。これらの研究は感情表現の抽出ルールや感情分類パターンの作成を常識知などから手動で行っており、抽出の自動化を目指す本研究とは方向性が異なる。

3. システム

本節では、本研究が感情表現抽出に関して提案するモデルとシステムについて述べる。本システムの特徴は、ブートストラップを用いることで、わずかな語やパターンから感情表現抽出のための辞書が自動的に構築できること、文字列の長さに依存しないパターンが半自動的に抽出できることが挙げられる。

3.1. 感情抽出のモデル

本研究では感情表現抽出に際し、テキスト中の感情表現のモデル化を行った。感情表現の抽出に際し、その処理を階層的に扱うため Lv.1~4 を定めた。Lv.1~Lv.4 は、それぞれが感情の表現形態を表しており、それぞれに対して抽出方法を提案した。

3.1.1. 感情表現 Lv.1

Lv.1 の感情表現は、「うれしい」「悲しい」など、感情語で直接表現されるものであり、感情語辞書を作成することで抽出が可能である。処理対象文字列は1語である場合が多く、抽出の難度も低い。

3.1.2. 感情表現 Lv.2

Lv.2 は、基本的には感情表現だと思われるが、場合によっては感情表現ではなくなるものである。例えば「好き」という感情表現があった場合、「X が好き」なら好意を表す感情表現と受けとることができるが、「物好き」「好き嫌い」といった熟語になると、感情表現であるとは言い難い。Lv.2 では、抽出する語が複合語や熟語の一部になっていないかどうかの判断が必要であ

る。これは感情語および前後の語も含めたパターンの辞書を作ることによって抽出が可能であると考えられる。処理対象文字列は2語以上であり、Lv.1 に比べて抽出難度は高い。

3.1.3. 感情表現 Lv.3

Lv.3 は、前後の文脈を判断しないと感情表現かどうか判断できない場合である。例えば次のようなテキストがあったとする。

私は、派手でうるさい音楽が大嫌いなのです。ですから、クラシック音楽や単一の楽器の音色を好んで聞きます。・・・(中略)・・・次に紹介しますこの CD は非常に派手で派手でうるさかったのです。

このようなテキストがあった場合、Lv.1 に相当するのが「大嫌い」や「好んで」であるが、ここで問題となるのは最後の文にある「この CD は非常に派手で派手で」である。この書き手は最初の文にて「派手でうるさい音楽が大嫌いなのです。」と宣言している。つまりこの場合、最初の文以降は「うるさい=大嫌い」「派手=大嫌い」という感情表現と考えることができる。このように、前後の文脈を見ることで初めて抽出できるのが Lv.3 である。処理対象文字列は複数の文であり、抽出難度は非常に高い。

3.1.4. 感情表現 Lv.4

最後の Lv.4 は、書き手そのものの性質を理解することではじめて感情表現であると理解できる場合である。極端な一例として、書き手が神経質な受験生であったときの場面を考える。その書き手が「今日帰り道のこと、路面の氷ですべて転んでしまった・・・」と記述していた場合、「すべて」という部分が「受験ですべて」につながるため、書き手としてはとても悲しい感情表現として記述している状況が考えられる。書き手その本人を実際に知ることはほぼ不可能であるが、少なくともその書き手が記述した他のテキストから「書き手が受験生である」ということを知ることはでき、それを手がかりにした感情表現の抽出が考えられる。これは、関連研究で紹介した Liu らの常識知による処理に近い。書き手自身の特性を理解したうえで、常識などによって推論する必要がある、感情表現の抽出難度は最も高いと言える。

以上のように、本研究では感情表現を4つの場合に分けて処理することを提案する。本稿では、その中でも Lv.1 と Lv.2 の処理のための辞書構築に関して述べる。また、表2に Lv.1~4 の関係を比較する。

表 2 モデルで設定した各レベルの比較

	処理対象文字列	抽出難度	抽出方法
Lv.1	ほぼ 1 語	低	感情語辞書
Lv.2	2 語以上	中低	パターン辞書
Lv.3	複数の文	中高	パターン辞書
Lv.4	テキストおよび書き手の情報全て	高	常識知

3.2. 感情表現の抽出と辞書構築

本研究では、感情表現の抽出と辞書構築をする手法としてブートストラップと Prefixspan, 係り受け解析を組み合わせた手法を提案する。ブートストラップは手がかりの少ない学習に有効であるが、学習する文字列長などに関するルールは決まっていない。そこで、Prefixspan およびブートストラップを組み合わせることで、少ない手がかりから文字列の長さ制限の無い学習を行った。その際、テキストは形態素解析器『茶筌』[14]を用いて基本形にしてから学習に用いている。感情語に係る頻出パターンを正しく抽出するため、パターン抽出時のみ係り受け解析機『南瓜』[15]による係り受け解析を組み合わせた。以降、本研究でのシステムを説明する。

3.2.1. ブートストラップ

本稿で用いたブートストラップに関して説明する。まず、感情語辞書とパターン辞書を定義する。感情語辞書には感情を表現する語（以下：感情語）が登録され、パターン辞書にはその前後に現れる語やパターンが登録される。感情語辞書を用いてパターン辞書を、パターン辞書を用いて感情語辞書を交互に作っていくことで辞書を構築していく。図 1 を用いて実例を示す。

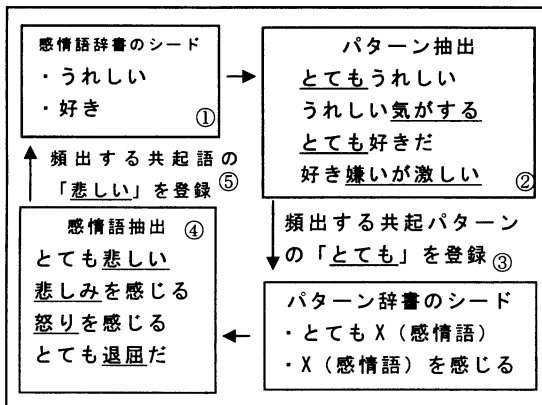


図 1 辞書間の相互関係

まず、感情語辞書にシードとなる

「うれしい」「好き」

が登録されているとする。次にこのシードを含むパターン

「とてもうれしい」「うれしい気がする」
「とても好きだ」「好き嫌いが激しい」

を抽出する。その中から、人手で頻出するパターンである

「とても X (感情語)」

をパターン辞書のシードへ新たに登録する。図 1 では、別の頻出パターンとして

「X (感情語) を感じる」

も登録している。次にパターン辞書のシードから、そのパターンが出現する

「とても悲しい」

「悲しみ（基本形：悲しい）を感じる」

「怒りを感じる」「とても退屈だ」

を抽出する。この中から、人手により頻出語である「悲しい」を感情語辞書に新たに登録する。このように、感情語辞書をシードとしてパターン辞書を、パターン辞書をシードとして感情語辞書を相互に繰り返し構築してゆく。辞書への登録が自動であるのに対し、シードは人手にて候補リストから選別し登録される。

3.2.2. Prefixspan

Prefixspan の手法について説明する。Prefixspan は Pei ら[16]によって提案された、シーケンシャルパターンマイニングの手法の一つで、深さ優先探索によって高頻度パターンを高速に抽出できる手法である。ワイルドカードを含むパターン抽出も可能である。特に抽出するパターンの文字列長に制限がない特徴に着目し、本研究では Prefixspan を採用した。図 2 に簡単な動作例を示す。id1~4 の文字列に対し、先頭に a を持つパターンを抽出したところ、7 種 12 個のパターンを抽出できた。本システムでは工藤らの開発した Prefixspan のプログラム[13]を利用した。

id	Sequence		
1	a	c	d
2	a	b	c
3	c	b	a
4	a	a	b
a, b, c, d それぞれの単独パターンを抽出 a:4 b:3 c:3 d:1			
id	Sequence (a)		
1		c	d
2		b	c
4		a	b
a に続くパターンを抽出 (a)a:1 (a)b:2 (a)c:2			
id	Sequence (a)(a)		
1			
2			
4			b
aa に続くパターンを抽出 (a)(a)b:1			
a を先頭に持つ全てのパターン a:4 (a)a:1 (a)b:2 (a)c:2 (a)(a)b:1 (a)cd:1 (a)bc:1			

図 2 Prefixspan の動作

3.3. 感情表現の抽出と辞書構築

以下、感情表現抽出と辞書構築の手順を示す。

1. 感情語辞書にシードとなる感情語を登録する
2. 形態素解析器『茶筌』を用いて基本形でのマッチングを行い、コーパス内で感情語を含む文を抽出する。
3. 抽出された全ての文に対して Prefixspan を用い、頻出パターンの抽出を行う。
4. 感情語を含む頻出パターン以外を削除する。結果、感情語を含む頻出パターンのみが残る。
5. (パターン抽出の場合のみ) 係り受け解析機『南瓜』を用いて係り受け解析を行い、感情語にかかる語のみを残す。
6. 最終的に残ったものが感情語に係る語とそのパターンになる。
7. 頻度情報をもとにして、感情語との共起パターンの上位をパターン辞書に登録する。このうち3件を人手で選び、シードとする。
8. パターン辞書と感情語辞書を入れ替えて 1~7 を行う。

以上の9手順を繰り返すことで、感情語辞書とパターン辞書を相互に構築した。

4. 評価実験

提案手法の有効性を評価するため、前節までに説明したシステムを用いて、実験を行った。

4.1. 実験方法

対象は杉田ら[11]によって Web 上から収集されたあるコミュニティに属する 326 サイト上の書評を中心とする 11648 テキストである。開始時のパターン辞書構築のための初期シードは「好き」「面白い」とした。3.2.1 節で定義した2つの辞書を3回ずつ構築した結果、感情語辞書には1129エントリ、パターン辞書には1022エントリが登録された。

4.2. 感情語辞書を用いた抽出精度

感情語辞書からランダムに3回50エントリを抜き出し、適切に感情表現が抽出できているかどうかを判定した。抽出された感情表現が適切かどうかの判定は、筆頭著者が行った。表3は判定の結果である。その結果、平均で0.35の精度を得た。3回の精度はほぼ同じ値を取っている。

4.3. 辞書組み合わせによる抽出精度

4.2 節とは別に感情語辞書からランダムに50エントリ、パターン辞書からもランダムに50エントリを抜き出し、組み合わせることで感情語表現として適切かどうかを判定した。前節と同様に、判定は3回行った。その結果として、平均0.28の精度を得た。感情語の判定と異なり、3回の精度には表3に示すようにばらつきが見られた。

表 3 抽出精度評価の結果

	精度の 最大値	精度の 最小値	精度の 平均値
感情語辞書 による評価	0.4	0.32	0.35
組み合わせ による評価	0.4	0.18	0.28

4.4. 考察

- 精度評価から得られた考察を示す。表4は判定の一部である。精度評価された感情語辞書の平均精度に比べ、パターン辞書の精度は低い。また、パターン辞書は精度の値にばらつきがある。感情語辞書の失敗では
- ・ 特定のサイトで集中的に使われていた特殊な表現を誤って学習
 - ・ 本のタイトルなどから誤った学習

などにより、誤った抽出や抽出漏れを起こしていた。パターン辞書の失敗では、

- ・ 頻出する感情語から多くのパターンを学習した結果、誤ったパターンを学習
- ・ ある特定の感情語しか取らないパターンをその他の感情語と組み合わせってしまった

などにより、意味をなさない事例が確認できた。組み合わせがうまくいった時には「やはり雲の上の上の人だよなあ」のように、感情語辞書だけでは抽出できない表現を抽出できた。これにより、パターンと感情語の組み合わせルールを作成し、誤った組み合わせを防ぐことが必要であるとわかった。

表 4 判定された例

	感情語 (基本形)	パターン (基本形)	判定
感情語辞書の判定	おもしろい		○
	た、という		×
	新鮮だ		○
	曖昧です		○
	せるためにどうするたいい		×
	あるます		×
	を認めるた		×
	やわらか		○
	この人の作品は上記の冊以外は何かを書いた		×
組み合わせによる判定	の量	X (感情語) です	×
	好きだ	X 人	○
	どうしていい	X だ	×
	例外	X だが	○
	いい	X ですか？	○
	人のせいだ	X 手	×
	に守られるている	いいかえるても X	×
	やはり雲の上の上の人だ	X よなあ	○
	読み込めが	ようだ	×

5. まとめ

本稿では、テキストに含まれる感情表現を半自動的に抽出するため、抽出のためのモデルを分類し、その一部のモデルに対処するためにブートストラップと Prefixspan を組み合わせる手法を提案した。主観的評価の結果、感情語の抽出に関しては一定の有効性を確認できたが、パターンとの組み合わせや抽出漏れなどに課題を残した。今後は感情表現であるかどうかの判断に対して複数判定者の一致度を見たり、抽出の精度向上、抽出ルールの再構築および再評価を行い、本提案手法の有効性を高めていきたい。

Lv.3 以降の感情表現に関しては、まだ有効な処理方法が考案されていない。こちらも合わせて検討し、テキストからの感情表現抽出のモデルとしても有効性を高めたい。

参考文献

- [1] 立石健二, 石黒義英, 福島俊一, インターネットからの評判情報検索, 人工知能学会誌, Vol.19, Num.3, pp.317-323, 2004.
- [2] Dave, K., Lawrence, S., Pennock, D.M. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. International World Wide Web Conference, Budapest, Hungary pp.519-528, 2003.
- [3] Pang, B., Lee, L., Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques, pp.79--86, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP2002), Philadelphia, July 2002.
- [4] Kobayashi, N., Inui, K., Matsumoto, Y., Collecting Evaluative Expressions for Opinion Extraction. Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04) pp.584-589, 2004.
- [5] Liu, H., Lieberman, H., Selker, T., A Model of Textual Affect Sensing using Real-World Knowledge. To Appear in Proceedings of IUI 2003 2003.
- [6] Nasukawa, T., Yi, J., Sentiment Analysis: Capturing Favorability Using Natural Language Processing. International Conference On Knowledge Capture: Proceedings of the international conference on Knowledge capture K-CAP '03 pp.70-77, 2003.
- [7] 館野昌一, 「お客様の声」に含まれるテキスト感性表現の抽出方法, 情報処理学会研究報告 自然言語処理 Vol.2003, Num.4, pp.105-112, 2003.
- [8] Turney, P., Littman, M., Measuring praise and criticism: Inference of semantic orientation from Association. ACM Transactions on Information Systems (TOIS) Vol.21, Num.4, pp.315-346, 2003.
- [9] Wilson, T., Wiebe, J., Hwa, R., Just how mad are you? Finding strong and weak opinion clauses. Proc 19th National Conference on Artificial Intelligence (AAAI-2004) 2004.
- [10] Riloff, E., Wiebe, J., Wilson T., Learning Subjective Nouns using Extraction Pattern Bootstrapping, Proceedings of the Seventh Conference on Natural Language Learning, Edmonton, Canada, May 2003.
- [11] 杉田茂樹, 江口浩二, 目録データベースと Web コンテンツの統合的利用方式, 情報処理学会研究報告 情報学基礎, Vol.2001, Num.20, pp.153-158, 2001.
- [12] 田中努, 徳久雅人, 村上仁一, 池原悟, 結合価値パターンへの情緒生起情報の付与, 言語処理学会第10回年次大会発表論文, pp.345-348, 2004.
- [13] 工藤拓, 山本薫, 坪井裕太, 松本裕治, 言語情報を利用したテキストマイニング, 情報処理学会研究報告 自然言語処理, Vol.2002, Num.20, pp.65-72, 2002.
- [14] 松本裕治, 北本啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸, 日本語形態素解析システム『茶釜』version2.2.9 使用説明書, 2004.
- [15] 工藤拓, 松本裕治, 日本語係り受け解析器『南瓜』version0.50, 2004.
- [16] Pei, J., Han, J., and et al, Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, In Proc. of International Conference of Data Engineering, pp.215-224, 2001.