社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (16)
2006－ＮＬ－175 (16)
2006／9／13

文構造の標準化によるKwicの拡張

秋山優† 深谷昌弘†† 大岩元††† 舘野昌一††††
†慶應義塾大学大学院政策メディア研究科
††慶應義塾大学総合政策学部
†††慶應義塾大学環境情報学部
††††富士ゼロックス株式会社研究本部(FXPALジャパン)
Email:maooon@sfc.keio.ac.jp

あらまし マーケティングリサーチでは，アンケート自由回答文等大量のテキストに記述された意見を，文脈を保持した形式で一覧集計表示する要素技術が求められている．しかし，こうした要求に対する標準技術であるKwicでは，同一の意味内容でも語順が異なる文や離れた文節間で係り受け関係が構成される文は，異なる文脈として表示されてしまう．この問題を解決するため，文構造を標準化した上で表示するシステムを提案する．提案システムによって，同一の構造を持つ文を類似意見として一覧集計表示可能であり，マーケティングリサーチにおける意見把握に有効であることを事例によって確認した．

Extending Kwic Concordance by standardization of sentence pattern

Yu Akiyama† Masahiro Fukaya†† Hajime Ohiwa††† Masakazu Tateno††††
†Graduate School of Media and Governance, Keio University
††Faculty of Policy Management, Keio University
†††Faculty of Environmental Information, Keio University
††††FXPAL Japan, Corporate Research Group, Fuji Xerox

Abstract In this paper, we have extended Kwic Concordance for marketing research. We think Kwic has two problems as a method for semantic survey on opinion information written on a large amount of text. One is that Kwic shows sentences which contain distant dependency as different context. Another is that Kwic also shows sentences which have different word order as different context. We propose the method to standardize and show sentence patterns, and we evaluate the method on marketing research.

1．はじめに

情報技術の発展によって，不特定多数の人々が記述した大量のテキストが，機械可読な形式で入手可能となった．例えば，Web等を介して大量のアンケート自由回答文が採集可能である．また，企業コールセンターでは，製品やサービス等に関する顧客の様々な意見がテキストとして大量に蓄積されている．

これらテキストデータには，従来の選択式アンケート等の意見調査では見出せない，書き手の自

由想起に基づく意見が含まれている．そのため，マーケティングリサーチでは，大量のテキストに記述された意見を定量的に把握するための技術が求められている[1,2,3,4,5].

このような要求に対する標準技術の一つとして，市販のテキストマイニングソフトは，Kwicによる原文表示機能を提供している．Kwic (Keyword-in-context) は，1959年IBMのH.P.Luhnによって考案された手法であり[7]，検索語句を中央に配置し，その前後文字列を検索語句の左右に文脈として表示する．検索語句の前後文字列を先行語順，あるいは後続後順にソート表示することができるため，Kwicは検索語句の様々な語られ方を把握する際に有効である．図1は，「ビールを美味しく飲むためのシチュエーション」に関する自由回答文を「風呂上り」をキーワードとして検索し，結果をKwicによって表示した例である．

ありきたりですが、　風呂上り　ですね。後
休日の前日に、　風呂上り　に、おいし
家族や友達と、　風呂上り　にわやわや
冷たくして、　風呂上り　に飲むとき
夏の暑いときに、　風呂上り　に飲むビー
仕事を終えて、　風呂上り　や食事前に

図1：Kwicによる原文表示例

しかし，テキストに記述された意見を定量的に把握しようとする際には，検索語句の前後文字列を文脈として表示するKwicは以下の点で不十分であると考える．

第一に，係り受け関係が離れた文節間で構成される文は，挿入をはさむために意味内容は類似していても異なる文脈として表示されてしまう．異なる文脈として表示された文はソートできず，類似意見として集計することが難しい．図2は，「風呂上りに一飲む」という係り受け関係を含む文をKwicにより表示した例である．大量の文の中に「風呂上りに一飲む」という係り受け関係が散在しているため，「風呂上りに一飲む」という係り受け関係を含む文を同一の文脈として表示できない．

お風呂上りに　飲む　事
暑い時期にお風呂上りに　飲む　時
風呂上りにくつろぎながら　飲む
風呂上りにひとりで　飲む　とき
り風呂上りにひとりでぐいっと　飲む
にしてお風呂上りにゆっくり　飲む

図2：離れた文節間で係り受け関係が構成される文をKwicによって表示した例

第二に，同一の意味内容であっても語順の異なる文は，異なる文脈として表示されてしまう（図3）．そのため，類似意見がソートできず集計が困難である．

風呂上り　に一人で飲む
一人で　風呂上り　に飲む

図3：語順の異なる文をKwicによって表示した例

このような問題を解決するために，本稿では文構造を標準化した上で表示するシステムを提案し，Kwicによる原文表示の拡張を行う．図4は，提案システムによる原文表示例である．

検索語句の前後文字列を文脈として表示するKwic表示の問題は，挿入や語順の異なりによって，同一の意味内容を持つ文が異なる文脈として表示されることであった．これを解決するため提案システムは，文の骨格となる係り受け関係のみを文脈として表示する．即ち，構文解析によって抽出された受け語（述語）及び受け語に係る複数の係り語（名詞句）を文の骨格となる構成要素とし，1行に1文の構成要素を文脈として表示する．挿入によって係り受け関係が離れた文節間で構成される文であっても，同一の構成要素を持つ文は同一文脈として表示可能である．また，助詞別に列を固定したセルに係り語を表示することによって，語順は異なっていても同一の構成要素を持つ文は同一文脈として表示可能である．

番号	は	が	を	に	で	その他	述語	分類	文
341				風呂上りに	一人で		飲む		夏の風呂上りに一人で飲む。
736				風呂上りに	一人で		飲む		一人で風呂上りに飲む
992				風呂上りに	一人で		飲む		一人で風呂上りに飲む
1087				風呂上りに一気に	一人で		飲む		風呂上りに一人で一気に飲む
303				風呂上りに		友達と	飲む		友達と風呂上りに飲む。
311				風呂上りに		友達と	飲む		スポーツ後の風呂上りに友達と飲む。
1554				お風呂上がりに		ぐっびと	飲む		お風呂上がりにぐっびと飲む
1659				風呂上りに		ごくごく	飲む		風呂上りにごくごく飲む
397				風呂上りに		冷やして	飲む		冷やして風呂上りに飲む。
1806			缶ビールを	風呂上りに			飲む		疲れたときの風呂上がりに缶ビールを飲む

図 4：提案システムによる表示例

提案システムでは意見を単文を単位として定義し，同一の構成要素を持つ単文を類似意見として一覧集計する．

なお，提案システムは，慶應義塾大学及び富士ゼロックス株式会社研究本部 FXPAL ジャパンが共同開発を行っているテクスト分析システムのモジュールである．同開発プロジェクトは，慶應義塾大学湘南藤沢キャンパスにおける 21 世紀 COE プログラム「日本・アジアにおける総合政策学先導拠点－ヒューマンセキュリティの基盤的研究を通して－」[8]の一環として，2003 年度より行われている．

2 節では提案システムの実現方法について述べる．3 節では提案システムを用いたマーケティングリサーチの事例を示す．4 節で評価を行い，5 節で課題について述べる．

2．アプローチ

前節では，Kwic による原文表示が大量テキストにおける意見の定量的な把握に不十分である理由を述べ，Kwic を拡張した原文表示システムを提案した．本節では，Kwic の問題点を解決するためのアプローチとして，文構造を標準化した上で表示する手法を述べる．はじめに，テキストから抽出すべき文構成要素及び抽出した要素の表示形式を述べ，機能要件を定義する．次に，機能要件を実現するためのシステム構成について述べる．

2－1．抽出すべき情報

本稿では，単文を単位として意見を表示する．また，表示された単文の意見性判定は人手によって行う．例えば，「料金が高いので，私は不便でも携帯を持ちません」という複文は，「料金が高い」「私は不便でも携帯を持ちません」という二つの単文から成り立つ．この時，提案システムはそれぞれの単文を意見の候補として表示し，意見性の判定は分析者が行う．

Kwic は検索語句の前後文字列を文脈として表示するため，挿入によって係り受け関係が離れた文節間で構成される文は，意味内容は類似していても異なる文脈として表示されてしまうことを述べた．この問題を解決するため，構文解析によって得られた係り受け情報を用いて，単文の骨格となる構成要素のみを文脈として抽出する．単文の骨格となる構成要素は，一つの述語と述語に係る 0 個以上の名詞句とする．また，名詞句とは，名詞と後続する助詞，及び名詞に係る連体修飾語から構成されるものとする．提案システムは，挿入の有無に関わらず，同一の構成要素をもつ単文を類似意見と見做す（図 5）．

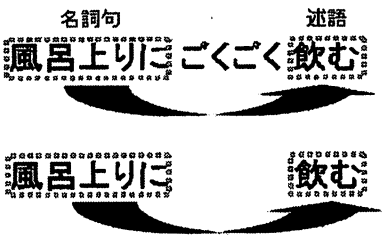

図5：単文を構成する要素

2－2．抽出した文構成要素の表示形式

Kwicによる表示では，同一の意味内容であっても語順の異なる文は，異なる文脈として表示されてしまう．この問題を解決するために，抽出した文の構成要素をどのように表示すべきかを述べる．

語順の異なりは，係り語となる複数の名詞句，即ち「名詞＋助詞」の位置が，比較的自由に入れ替わることによって生じる．そこで，構文解析によって抽出された係り語となる名詞句を，「は，が，を，に，で」の各助詞別に列を固定したセルに表示する（図4）．

係り語を助詞別に列を固定したセルに表示することで，語順に関わらず，同一構成要素を持つ単文を同一文脈として表示可能である．係り語及び受け語を頻度順にソート表示することによって，同一の構成要素を持つ文を類似意見として一覧集計可能である．

2－3．機能要件

提案する意見表示システムは，大量テキストに記述された単文を意見の単位とし，構文解析によって得られた構文情報を用いて単文の骨格となる構成要素を文脈として表示する．これまでの議論を踏まえて，提案システムに必要とされる主な機能要件を以下に示す．

(1) テキストから単文集合を検索する機能

(2) 検索の結果抽出された単文集合について，文構造を標準化した上で表示する機能

2－4．システム構成

図6は，上記機能の実現を目的としたシステム構成である．

提案システムでは，単文の骨格となる構成要素を抽出するために，構文解析を利用する．前処理として，CSV形式のテキストデータを入力として形態素解析及び構文解析を行う．得られた構文情報を利用し，文の骨格となる係り受け関係を構成する語句をデータベースにテーブルとして保存する．得られたテーブルを対象として検索を行い，検索によって抽出された単文の構成要素を一覧集計表示する．また，出力結果をExcelに出力することが可能である．

以下にシステム構成の詳細について述べる．

2－4－1．言語解析部

言語解析部では，CSV形式のテキストデータを入力として，形態素解析及び構文解析を行う．構文解析によって得られた構文情報を用いて，各単文の骨格を構成する語句を，データベースにテーブルとして保存する．言語解析部による処理は，前処理としてはじめに一度行えばよい．

構文解析には，米国 Xerox Corporation の Xerox Research Centre Europe (XRCE)によって開発された XIP (Xerox Incremental Parser) を利用した．

2－3－2．検索部

検索部では，DBに保存された言語解析部の処理結果を対象とした係り受け検索によって，任意の単文の構成要素を抽出する．

係り受け検索とは，構文解析によって得られる構文情報を利用した検索方式であり，「ある語がある語に係る」という語句間の係り受け関係を検索キーとして利用する．日本語を対象とした検索システムでは，従来検索語句が文書中に出現するか

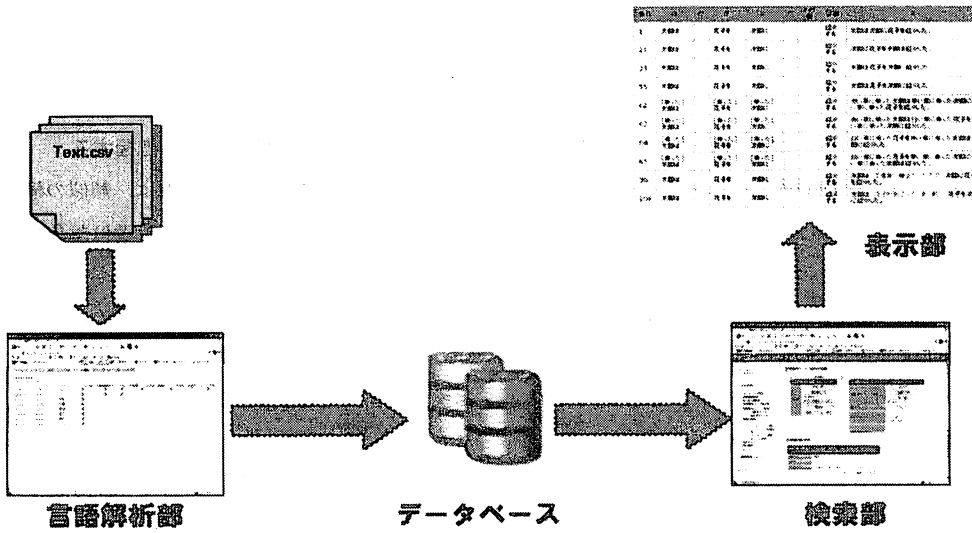

図6：システム構成

否かに関するブール演算を利用する方式が多く用いられてきた．係り受け検索は，構文解析等の自然言語処理技術の発展に伴い，より高精度な検索方式として情報検索や質問応答分野において提案されている．[6]

係り受け検索によって，ある名詞がどのような述語に係るか，あるいは，ある述語がどのような名詞を受けるかを検索可能である．例えば，商品のイメージ，評価や，要望表現等の対象を検索することが可能であり，マーケティングリサーチにおける意見調査に有効であると考える．

検索条件として指定可能な要素を以下に示す．

(1) 係り語

係り語として，名詞を指定可能である．何も指定しない場合，wildcard となる．また，指定した名詞と同一文節内にある「は，が，を，に，で」の各助詞を，OR 条件で組み合わせて指定可能である．

(2) 受け語

受け語として，述語を指定可能である．何も指定しない場合の効果は，係り語における場合と同様である．語句を指定せず，動詞，形容詞，形容動詞，名詞の各品詞を，OR 条件で組み合わせて指定することが可能である．

2－3－3．表示部

表示部は，係り受け検索の結果抽出された単文の骨格となる語句を，一覧表示する．

一覧表示は，行毎に一文の骨格となる構成要素を表示する．各行内では，係り語となる名詞句を「は，が，を，に，で」の助詞別に配置し，受けとなる述語，原文をその右側に表示する．上記以外の助詞を伴う係り語は，「その他」の項に表示される．また，係り語を連体修飾する修飾語は，参考情報として同一セル内に表示させることができる．述語は，終止形もしくは原文における活用形を選択表示可能である．

提案システムでは，係り語及び受け語を頻度順にソート表示することによって，類似の骨格要素を持つ単文が一箇所に集められて表示される．表示結果を閲覧することによって，類似の構成要素を持つ単文を類似意見として把握しやすいと考えられる．また，類似の構成要素を持つ文を集計表

示することも可能である．

3．解析例

提案システムを用いた解析例を以下に示す．本例は，「ビール，発泡酒等飲料に関する調査」として行われたアンケート調査の一部である．ここでは，「あなたがビールを飲むときに，あなたにとって最も美味しい飲み方とはどのような飲み方ですか．美味しく飲めると思う場面をできるだけ具体的にわかるようにお書き下さい」という設問に対する自由回答文 1,545 件を解析対象とした．

システムの出力した一覧集計表示の一部を図 7，8 に示す．図 7 は，受け語として「飲む」を指定して検索を行った結果を一覧表示した一部であり，図 8 はその集計表示である．また，出力結果について人手によるアフターコーディングを行い，Excel を用いてグラフ化した例を図 9 に示す．なお，予め辞書によって同義語，類義語の統制を行っている．

4．Kwic との比較による主観評価

図 10 は，前節で対象とした自由回答文を対象に，「飲む」を検索語として検索を行った結果を Kwic によって表示した一部である．

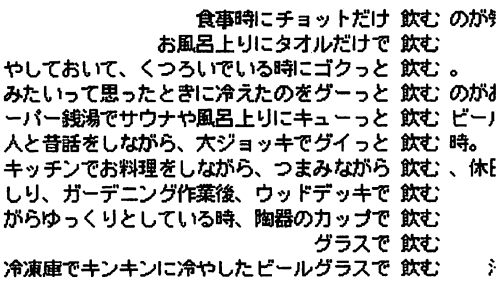

図 10：Kwic による表示例

マーケティングリサーチにおける意見把握では，分析者が様々な意見を一定の分類軸に沿って分類するアフターコーディングが重要である．例えば，「仕事帰りに飲む」「運動後に飲む」といった意見を「疲れた時に飲む」というカテゴリに分類し，各カテゴリ毎の意見数を集計，グラフ化することによって意見の全体的な傾向を把握する．

提案システムによる出力結果は，類似の構造を持つ文を集計表示するため，分析者は対象テキストにおいてどのような意見が何件述べられているかを把握することができる．そのため，アフターコーディングを行う際どのような分類軸を立てるべきかの判断を各意見の数量に基づいて行うことができる．一方，Kwic による表示では，意見を定量的に把握することができないため，分類軸の立て方が恣意的になりやすいと考えられる．

また，意見毎の件数が集計可能であるため，人手によるアフターコーディング作業を効率的に行うことが可能である．図 14 に示した分類結果を作成するのに要した時間は約 1 時間である．

5．おわりに

本稿では，テキストに記述された書き手意見の定量的な把握を目的として，構文解析を用いて文構造を標準化した上で表示するシステムを提案した．提案システムによって，類似の構造を持つ文を類似意見として一覧集計表示可能であり，マーケティングリサーチにおける意見把握に有効であることを Kwic との比較によって確認した．

今後の課題として，まず，意見性の判定が挙げられる．通常，テキストには伝聞，疑問，条件等の表現が含まれるが，提案システムでは意見性の判定を人手に委ねている．これは，意見性判定には文脈を考慮した意味解釈が必要と考えられるためである．

また，抽出精度向上も今後の課題である．提案システムを用いた解析例では，構文解析が適切になされず，本来除外されるはずの挿入節の一部が表示されてしまう等の現象が見られた．

6．参考文献

[1] 乾裕子，内元清貴，村田真樹，井佐原均，"文脈表現に着目した自由回答アンケートの分類"，情報処理学会研究報告，98-NL-128, p.181-188(1998)

[2] 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, No 3, pp.317-327, 2004

[3] 立石健二，福島俊一，小林のぞみ，高橋哲朗，藤田篤，乾健太郎，松本裕二，" Web 文書集合からの意見情報抽出と着眼点に基づく要約生成"，情報処理学会研究報告，04-NL-163, pp. 1-8(2004)

[4] 舘野昌一，" テクスト感性表現の抽出によるお客様の声の分析方法"，第 4 回日本感性工学会大会予稿集(2002)

[5] 舘野昌一，"「お客様の声」に含まれるテクスト感性表現の抽出方法"，第 9 回年次大会，言語処理学会(2003)

[6] 竹内淳平，辻井潤一，"係り受け関係と言い換え関係を用いた柔軟な日本語検索"，言語処理学会第 11 回年次大会発表論文集，pp.568-571(2005)

[7]Luhn, H. P., " Keyword-in-context index for technical literature(KWIC index) ", Yorktown Heights:IBM, 1959

[8] http://coe21-policy.sfc.keio.ac.jp/ja

番号	は	が	を	に	で	その他	述部	分類	文
341				[夏の]風呂上りに	一人で		飲む		夏の風呂上りに一人で飲む。
736				風呂上りに	一人で		飲む		一人で風呂上りに飲む
992				風呂上りに	一人で		飲む		一人で風呂上りに飲む
1087				風呂上りに|一気に	一人で		飲む		風呂上りに一人で一気に飲む
303				風呂上りに		友達と	飲む		友達と風呂上りに飲む。
311				[スポーツ後の]風呂上りに		友達と	飲む		スポーツ後の風呂上りに友達と飲む。
1554				お風呂上がりに		ぐっひと	飲む		お風呂上がりにぐっひと飲む
1659				風呂上りに		ごくごく	飲む		風呂上りにごくごく飲む
397				風呂上りに		冷やして	飲む		冷やして風呂上りに飲む。
1806			缶ビールを	風呂上りに			飲む		疲れたときの風呂上がりに缶ビールを飲む

図 7：「飲む」を受け語として検索した結果の一覧表示

	係り	受け	頻度	グラフ
1	風呂上がりに	飲む	137	
2	仕事帰りに	飲む	92	
3	家族と一緒に	飲む	88	
4	食事と一緒に	飲む	78	
5	夏に	飲む	75	
6	グラスで	飲む	67	
7	一人で	飲む	62	
8	暑い時に	飲む	55	
9	仕事が	終わって飲む	34	
10	運動後に	飲む	25	

図 8：集計表示例

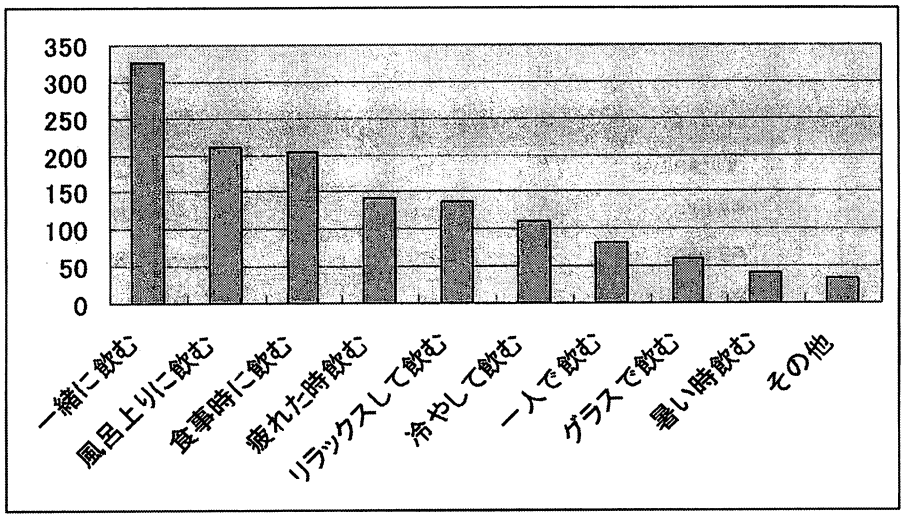

図 9：出力結果をアフターコーディングしグラフ化した例