意味解析が開く自然語処理の世界

青山学院大学教授　原田実

Tel:042-759-6321, Email:harada@it.aoyama.ac.jp

これまでの自然言語処理は、形態素解析、係り受け解析までが実用域に達しています。今回我々が開発した意味解析システムによって、自動要約、質問応答、内容検索、文分類、テキストマイングなどが意味を理解した高度なものに進化した現状を報告します。

１．意味解析システム SAGE

今回われわれが実現した意味解析システム SAGE は、juman/knp を用いた形態素解析・係り受け解析

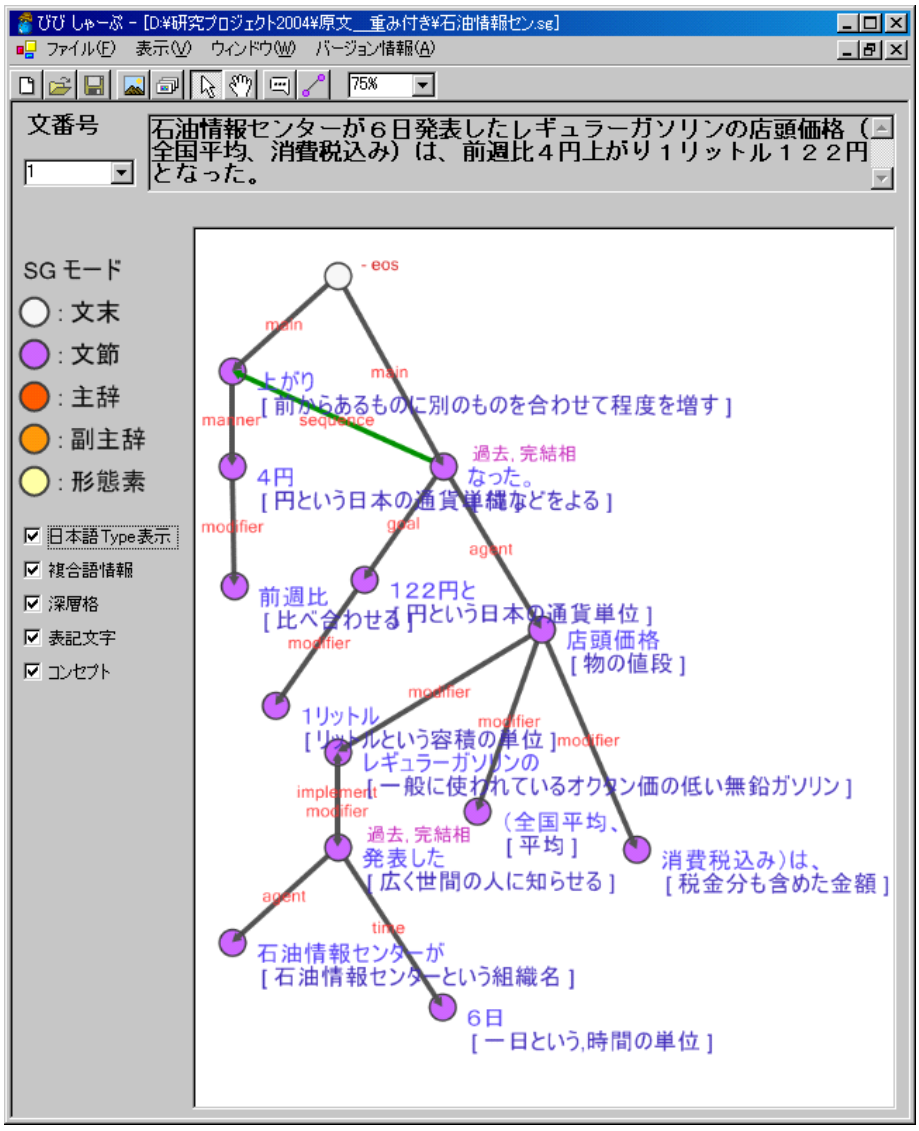


図１　上部窓枠内の文を意味解析した結果の意味グラフを図示した物。例えば、左下で文節「石油情報センター」は文節「発表した」の agent（行為の主体）的役割を演じていることが分かる。また、各文節には、文節内で中心となる形態素（主辞といわれる）の語意の説明が[]内に表記されている。

の後を受け、各語の意味を明らかにし、また係り受け関係にある（文節の代表）語間の役割的関係を明らかにします。SAGE は、語の意味を調べるのに EDR 電子辞書を用います。SAGE による意味解析結果は、文節や形態素ごとにそれらの意味や品詞や他の文節との役割的関係（深層格という）などを保持したリストの集合として表現されますが、文節を頂点、文節間の深層格を辺と考えたグラフを意味グラフと読んでいます。先の例文の意味グラフを、ビジュアルツール vivi によって図示すると図１のようになります。SAGE はこれらを独自の辞書展開法によって、高速（SAGE 全体：約 80～170 字/秒、juman/knp による形態素解析・係り受け解析部:90～180 字/秒、SAGE 本体の意味解析部のみ）:1000～1600 字/秒）かつ高精度（人手による正解に比べて語意 94%、深層格 87%）に行います。

このように文を構成する語の意味が分かると、例えば企業が持っている膨大な文書データから良質なセールス知識やマーケッティング知識などが抽出できるようになります。これはテキストマイニングといわれる技術で、文書データからの知識の発掘として注目を浴びています。しかし、従来のテキストマイニングでは、文書データに形態素解析と係り受け解析を行い、その結果の単語群をその表記や読みを伴って関係データベースなどに登録し、このデータベースに対して得たい情報をキーワードで指定して、表記の一致を条件にキーワードを含む文を抽出するなどをしていました。しかし、同一の概念を表す語はたくさんあり多様な表現で記述された実際の文書から思い通りの知識を検索することは大変です。これに対し、意味解析を行い、語に語意や係り先の役割を表す深層格を伴ってデータベースに登録すると、システムが文書中の各語と指定されたキーワードとの語意ベースの類似度を計算して、一定以上の類似度を持つ語を含む文を一括して抽出するので、キーワードの選択に迷うことなく求めている文を抽出することができます。

２．自然語仕様からのオブジェクト指向分析 CAMEO

意味解析技術を使えば、文章表現された内容を別の形式に変換するプログラムを簡単に作成することができます。例えば、文章に記されたソフトウエアに対する要求仕様書をもとに、オブジェクト指向分析を行い、図２に示すような設計図を自動的に作成できます。ここでは、プログラム内にオブジェクト指向分析の知識として、「教師や医者や作家や会社員などといった人の役割はクラスとして設計する」などのルールが蓄えられています。このような知識をプログラミングする場合、従来の形態素解析をベースにしたシステムでは、入力した文の各語が人の役割を表すものであるかを検査するために、「教師」や「医者」や「作家」や「会社員」などの人の役割をすべて列挙する必要があり、結果としてルールの数が非常に多くなり、維持管理も難しくなるという問題がありました。一方、意味解析された結果を入力すると各語には語意が付いているので、これらが上位概念として“ 444df2; 職業,肩書,役割で限定した人間 ”を持つかどうかを EDR の概念体系辞書で検査するだけで、その語が人の役割を表わす語かどうかを判断するこができます。言い換えれば、無数のルールが１つのルールと概念体系辞書に置き換わっているのです。これは、自然語文を入力して何らかの知的処理を行うプログラムの簡略化という点で非常に大きな意味を持ちます。

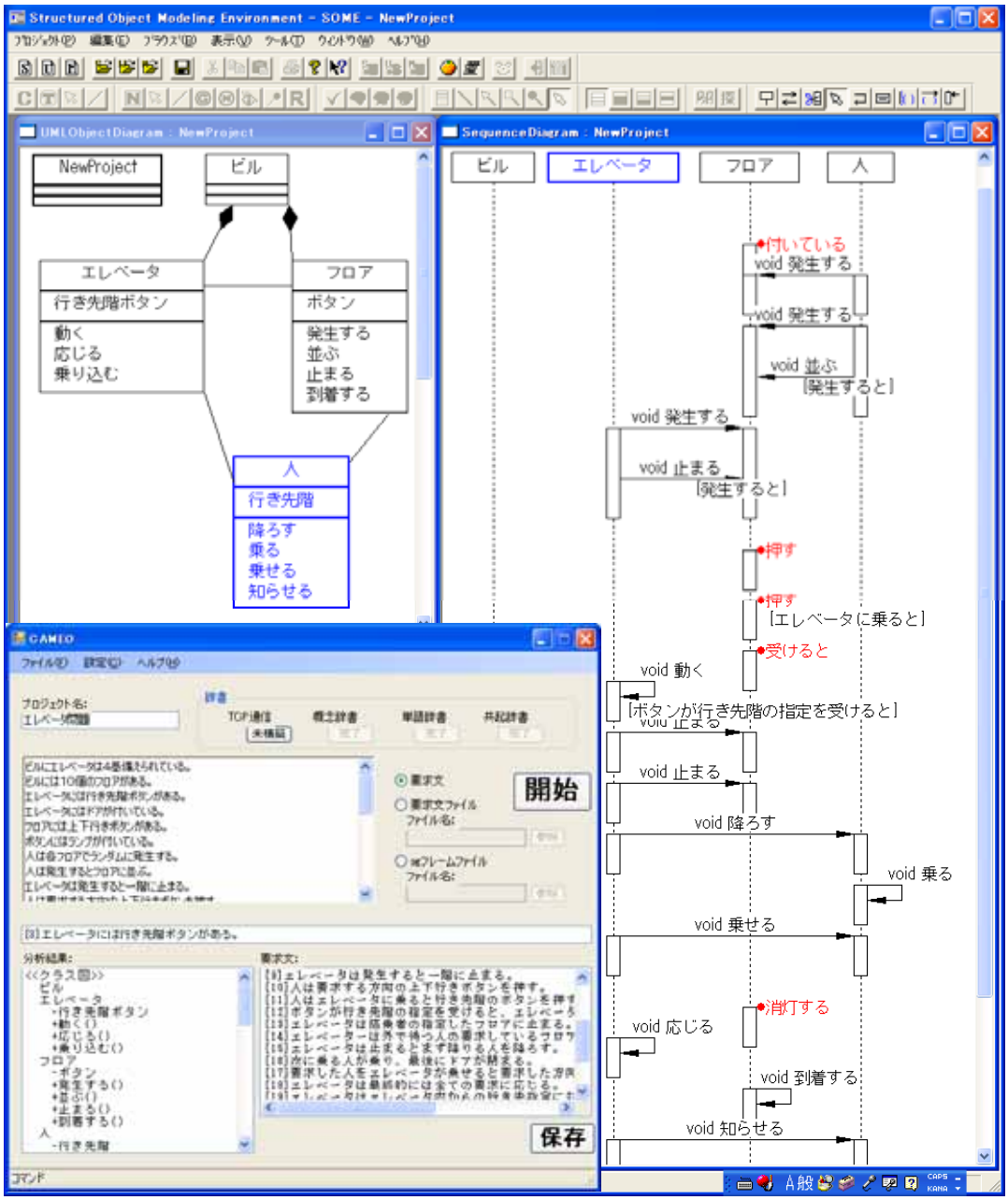


図２　左下の文章を意味解析後、オブジェクト指向分析ツール cameo が内蔵した分析知識に従って、右のシーケンス図や左上のクラス図を生成した。

３．自動要約システム ABISYS

市販の Word などの文書作成ツールに付随している要約機能は、形態素解析や係り受け解析に基づいた重要文抽出を行う機能であり、文章を構成する文の中から重要な語を含む文を選んで色づけするだけです。文を選定する尺度が甘く取り出された文よりも取り出されなかった文の方が重要であることも多く、また要約文が原文と同じなので短くもならないという欠点があり、実際に利用されることはほとんどありませんでした。我々が開発した自動要約システム ABISYS は、図３に示すように意味解析結果を用いているので良質な要約文を作成できます。

自動要約においては、要約文のタネとなる重要語の選定が成否の鍵を握ります。この際従来は、語の出現頻度を計算して頻出語を重要語としていました。しかし、通常の文章では同じ概念を表すのに同じ

語を用いずに別の表記で表現する場合が多いので同一表記の語を数えるだけでは見落としにつながります。また、「紅茶」、「茶」、「ウーロン茶」などが文章に出現すれば、これらをまとめて「飲み物」が話題の中心であることを判断しなければなりません。このような判断を行うためには、同じような概念を表すものは１まとまりとして頻度を計算しなければなりません。このような、語の意味的類似度の計算は、各語に与えられている語意をベースに概念体系辞書を検索することで実現できます。

また要約文の生成においては、重要語を種にできるだけ短い文を生成する必要があります。このためには、文中で各文節が係り先にとってどの程度必要な文節かを判断する必要があります。これは、意味解析によって係り先からみて各係り元に割り当てられた深層格を基に判断することができます。

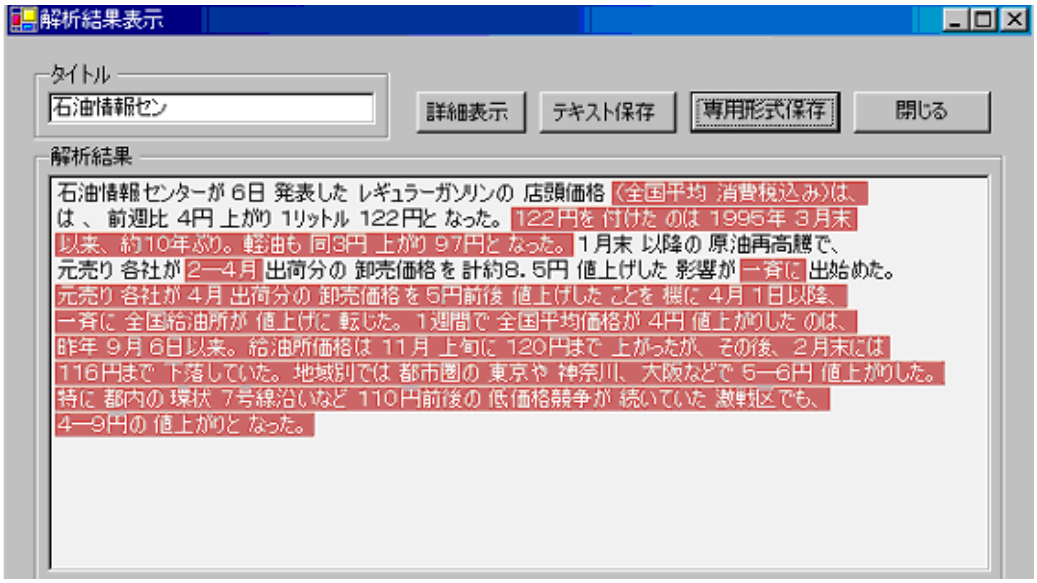


図３　画面内の全文を SAGE が意味解析し、その結果を基に ABISYS が要約として生成した要約文が白地に黒字で記されている。赤地に白字は要約としては不要とされた部分。

４．質問応答システム METIS

文を意味解析すると、語の語意や語間の深層格をもとに、２つの文の類似度を計算することができます。具体的には、文を意味解析した結果の意味グラフ上の共通部分グラフの大きさをベースに類似度を計算します。共通部分グラフとは、２つのグラフ上の対応する頂点同士が類似した語意を持つ語を表し、さらにそれらの間の深層格も類似している頂点と深層格からなるグラフのことで、それが大きいほど２つのグラフはより広範囲に類似していると言えます。類似度を効率的に求めるアルゴリズムを考案し、質問文に対する意味グラフとインターネットから検索された知識文に対する意味グラフとの類似度を求めることによって、最大の類似度を持つ知識文における質問箇所に対応する頂点が表す語を質問に対する回答として生成する図４に示すよう質問応答システム METIS を作成しました。クイズミリオネアの 100 問に対してインターネットから回答を求める実験では、3 位回答までの正解率は 74%で、正解を含む知識文が得られていれば 80%の精度で正解を抽出することができます。

５．セマンティックテキストマイニング STM

自由記述式のアンケートは選択式のアンケートに比べ、回答者の自由な意見を集約できるなどのメリットがあります。しかし、大量のテキストデータを人手で分類し、分析するには多くの時間と人材の確

保が必要です。そのため、近年注目を浴びているのがテキストマイニングです。既存のテキストマイニングでは形態素の TF/IDF 値のコサイン距離による類似性を基にした分類を中心とした解析が行われていますが、結果として語意や語間の関係が把握できない、「何がどうだ」「何がどうした」などの複数の語からなる関係を把握できない、表現の揺れによる差異を吸収できないといった問題が顕在化しています。このような背景を踏まえ、我々は文が表す意味的内容の類似性をベースにしたテキストマイニングシステム STM を開発しました。STM では、図 5 に示すように、日本文を意味グラフに展開し、2 つの意味グラフの対応するノード同士の概念的な類似度やノード間の深層格の類似度をベースに、類似部分グラフの大きさで 2 文の類似度を計測することによって、表現が異なっていても同様な趣旨をもつ文を同意見として集約し分類します。さらに、図 6 に示すコレスポンデンス分析では、アンケートデータ(文

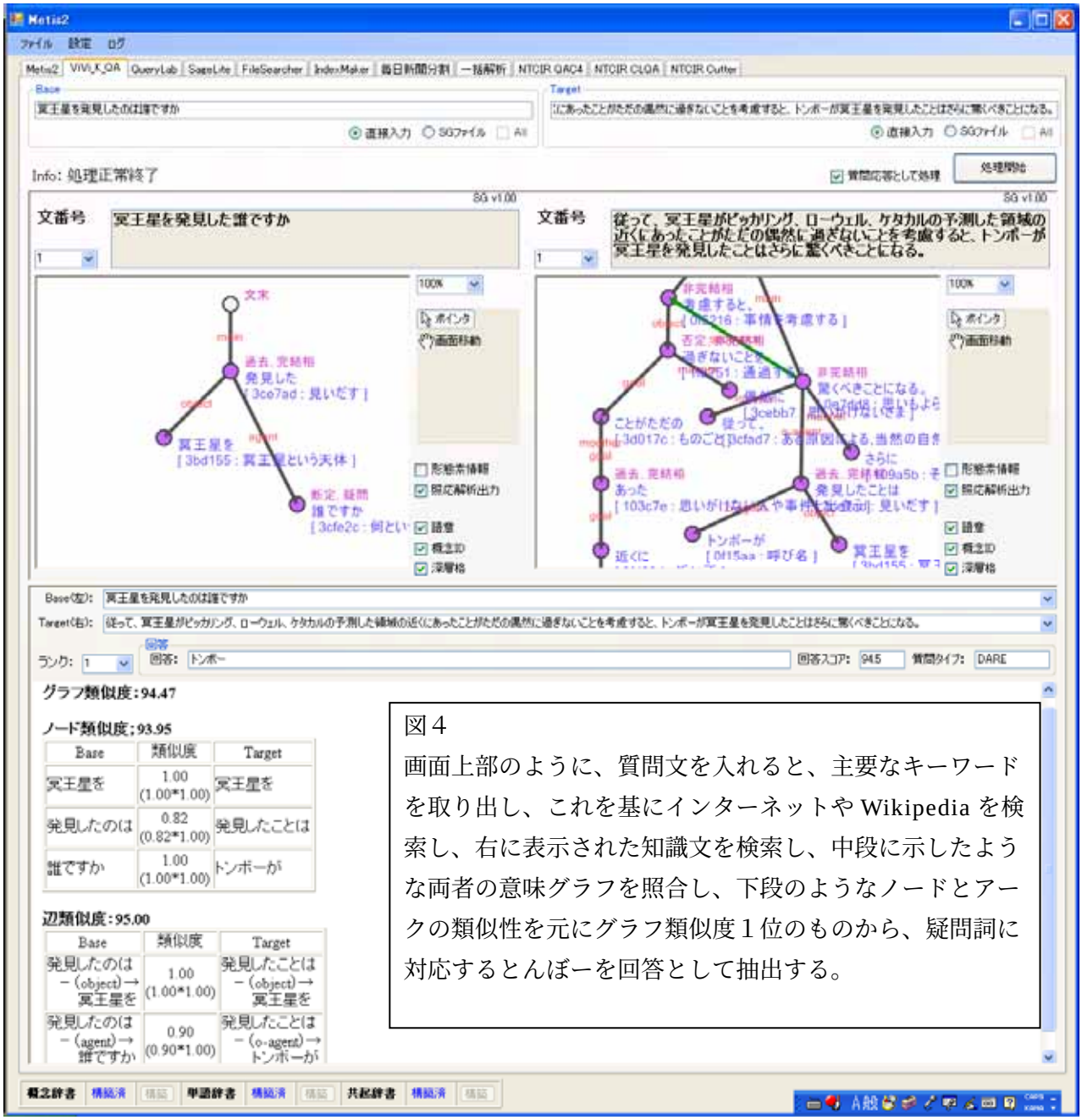


図 4
画面上部のように、質問文を入れると、主要なキーワードを取り出し、これを基にインターネットや Wikipedia を検索し、右に表示された知識文を検索し、中段に示したような両者の意味グラフを照合し、下段のようなノードとアークの類似性を元にグラフ類似度 1 位のものから、疑問詞に対応するとんぼーを回答として抽出する。

や句)のクラスタリング結果から、クラスタの要約文と要素数を回答者の属性ごとに分類したものをクロス集計表としてまとめ、それを基にポジショニングマップを作成します。回答者の属性は属性リストの項目を選択することで指定され、この項目に合ったデータ数がカウントされます。その後、作成されたクロス集計表のマトリックスよりポジショニングマップにおける回答者の属性および意見(クラスタの要約文)のプロット座標を算出し、図６に示すように表示します。

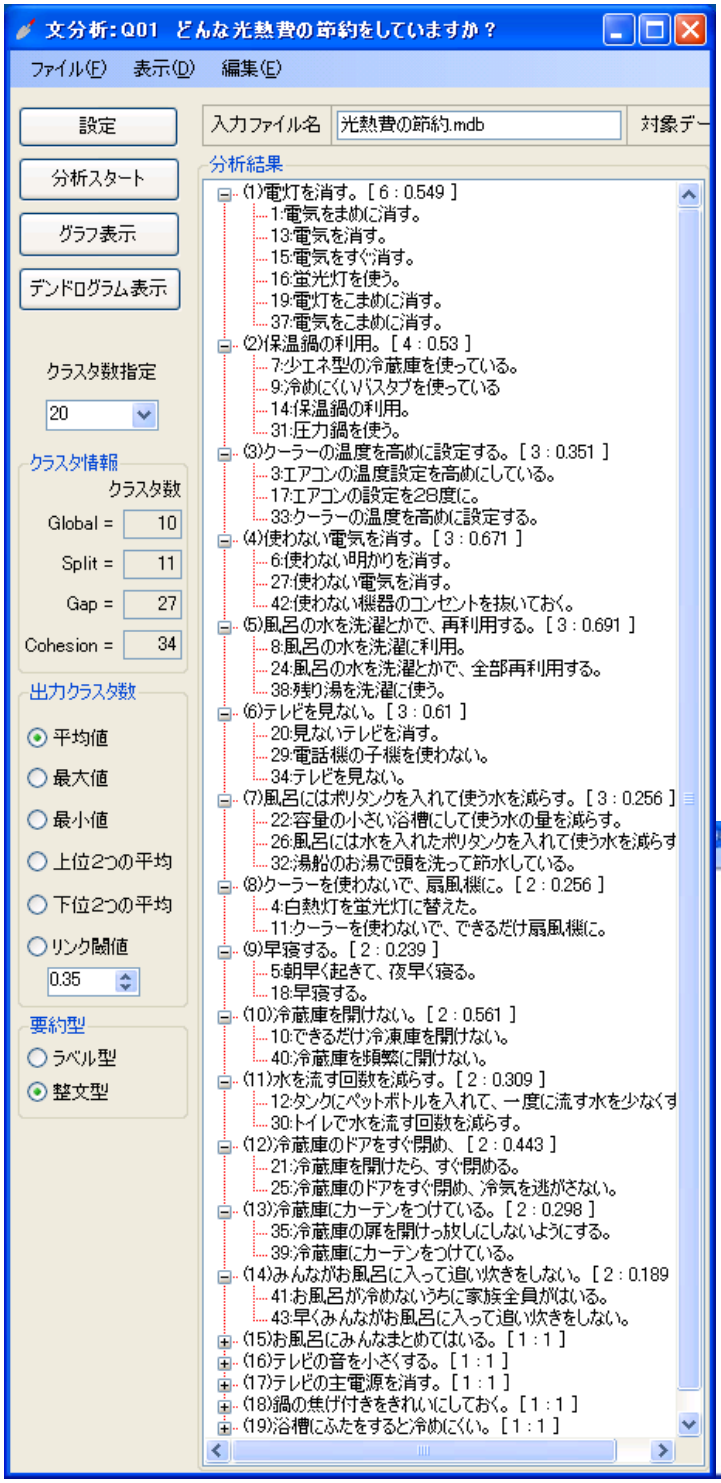
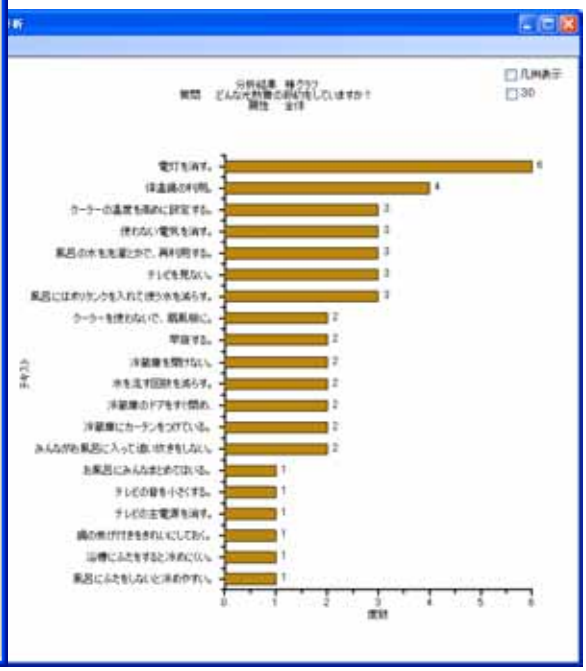


図５

光熱費の節約に対するアンケート 45 件を STM が 20 個のクラスタに分類している。意味をベースに、「何がどうだ、何がどうする」の観点から分類するので、概念の類似性を元に表層的に違うものでも同じ意味を表す意見を 1 つのクラスタにまとめている（「電気」と「蛍光灯」と「電灯」、「風呂の水」と「残り湯」、「見ないテレビを消す」と「テレビを見ない」）、一方細かな違いが正確に認識されるので同じ語でも使われ方やムードが違うと異なるクラスタに分類される（「冷蔵庫を開けたら、すぐ閉める」と「冷蔵庫の扉を開けっ放しにしないようにする」、「蛍光灯を使う」と「白熱灯を蛍光灯に替えた」）。



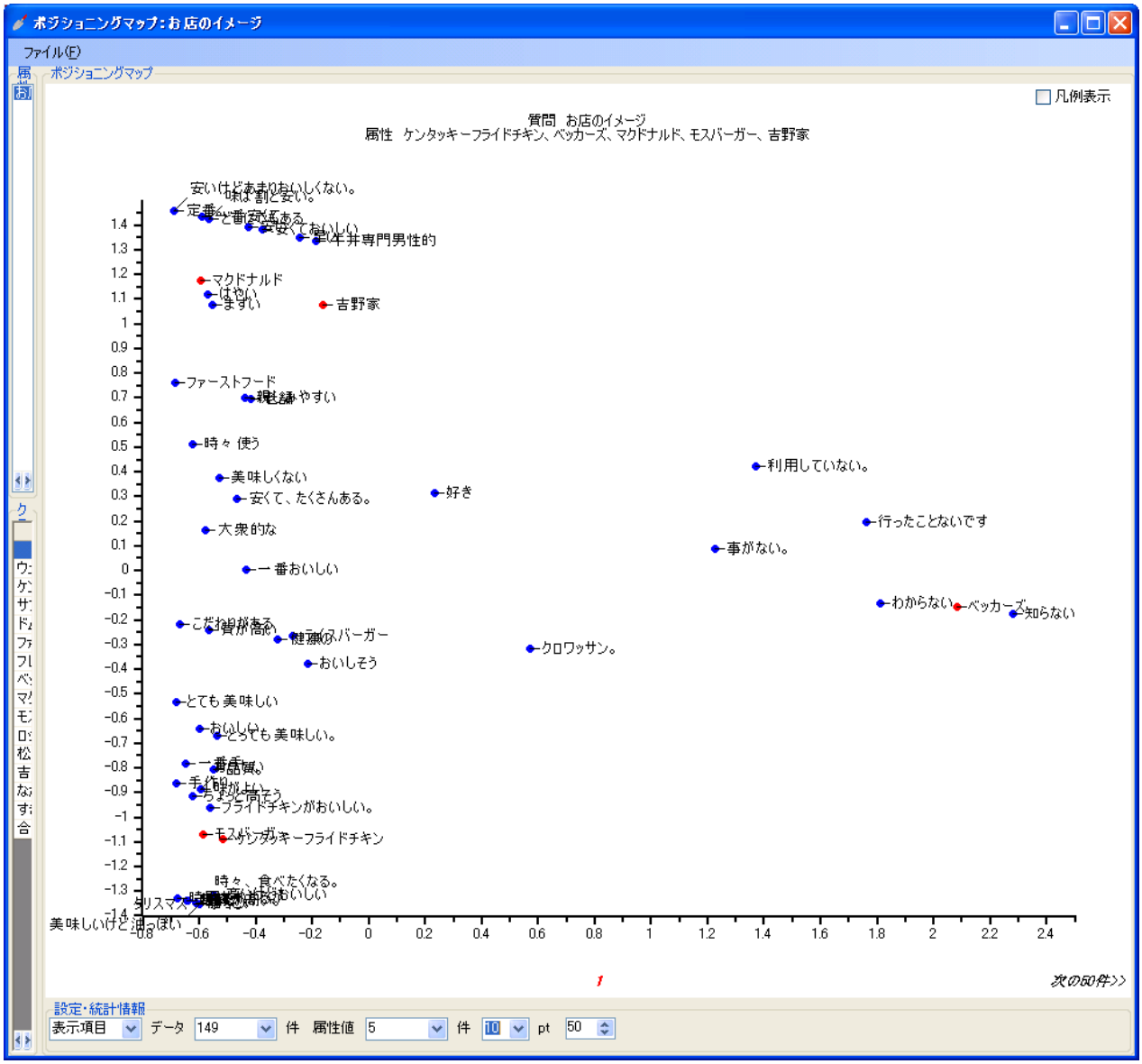


図6

ファーストフード転移対する顧客のイメージが 2 次元の座標上にプロットされている。赤が属性で、青がアンケートのクラスタの要約であり、各属性に多い意見が距離的に近いところに配置されている。