５Ｗ１Ｈ分類・ナビゲーションによる情報活用プラットフォーム

奥村 明俊　　池田 崇博　　村木 一至
NEC C&C メディア研究所

キーワード検索では、キーワード間に論理的なつながりがない情報も検索され、ノイズとなることが多い。そのためキーワード検索によって、ある出来事に至るまでの経緯、出来事に関する比較、情報全体の鳥瞰情報を有効に抽出することは困難である。本稿では、テキストに含まれるWho（だれが）・When（いつ）・Where（どこで）・What（なにを）・Why（なぜ）・How（どうした）という5W1H情報に着目し、キーワードの文中で果たす役割の観点にしたがって、ユーザーをナビゲートする5W1H分類・ナビゲーションを提案する。5W1H分類ナビゲーションは、エピソード抽出、多視点分類、情報鳥瞰の３つの機能を提供することによって、上述の問題を解決する。新聞記事8000件とセールスレポート2500件を対象として、5W1H分類・ナビゲーションを適用しその有効性を確認した。

Information sharing platform based on 5W1H clustering and navigation

Akitoshi Okumura, Takahiro Ikeda,　Kazunori Muraki
NEC C&C Media Research Laboratories

Keyword-based information retrieval techniques are difficult to provide process information about an event, comparative information about an event, and overall trend information about all events. This paper proposes an information navigation model by using 5W1H (when, who, what, where, why, how) information. 5W1H navigation model provides three functions: episodic information extraction, multiple view-points clustering, and overall information view. The model was effective when it was used for 8000 newspaper articles and 2500 sales reports..

1 はじめに

インターネット、イントラネットが急速に拡大し、情報洪水と呼ばれる程、多くの情報が氾濫している。数百ギガの大規模テキストを超高速に検索する技術が実現されたが、多くの業務において情報検索は手段であって目的ではない [1]。特に、オフィス業務においては、情報を検索しながら頭を整理し、アイデアを練ることが必要となる。現在の検索技術は、目的とは独立に情報を提供するものであり、目的を達成するためには多くのギャップが残されている。このギャップを埋めるためには、目的に最も合致した情報の選別と、やりとりをしながら知りたい情報の核心に迫るナビゲーション技術が必要である。

そこで、ユーザの目的に直結する情報集配信サービスとして、ユーザーの業務内容・技術分野等を、組織体系・技術体系等の複数の観点から構造化した複合オントロジを利用して、情報を選別し、配信するサービス MI-IDAS (Multi-Indexing Information Dissemination & Acquisition Service) の実現を目指している [2, 3]。また、情報分類・ナビゲーション技術として、テキストに含まれる Who (だれが)・When (いつ)・Where (どこで)・What (なにを)・Why (なぜ)・How (どうした) に対応する 5W1H 情報に着目し、5W1H の観点にしたがって、情報を分類・整理することで、混然とした情報の中でユーザーをナビゲートする 5W1H 分類・ナビゲーションを開発している [4]。

5W1H は、日常の出来事を理解するためのキーとなっている概念であり、5W1H 情報により、出来事の内容の核心部分が表現される。このため、5W1H 情報を利用することで、機能的で分かりやすい形で情報を提供することができる。

本報告では、まず、オフィス業務における情報検索の課題について述べる。つぎに、5W1H 情報を用いてこれらの課題を解決する 5W1H 分類・ナビゲーションについて説明する。そして、5W1H 情報抽出の手法について述べ、さらに、5W1H 分類・ナビゲーションを用いた情報活用プラットフォームを構築し、新聞記事とセールスレポートに、5W1H 分類・ナビゲーションを適用した例を示す。

2 オフィス業務における情報検索の課題

オフィス業務において、各種報告書、提案書、帳票などを作成する場合、参考となる文書を引用したり、データベースや WWW などの情報源から新たな情報を獲得するなど、多くの場合なんらかの検索行為を伴う。この場合、検索行為は手段であって、目的は文書作成である。

キーワード検索やフィルタリングサービスなど情報検索手法によって、ユーザーは、必要な情報を選択的に獲得することができるが、文書作成という目的からすると断片的な情報の集合である。目的遂行のためには、2 次的な検索要求が発生し、その都度、キーワードを生成して情報獲得と情報整理を行なうこととなる。オフィス文書作成を目的とした場合、1 次情報獲得の結果を基に、2 次情報獲得として、以下の 3 種類の情報展開が行なわれる。

時間的経緯: 縦方向への展開

獲得した情報が、発生するに至る時間的な経緯情報を検索する。例えば、「A 社が～ギガのメモリを開発した」という情報に対して、A 社が今までメモリの開発に関してどのような時期にどのような技術を開発してきたのかという、いわば、1 次情報からの垂直方向、縦方向への展開がある。

類似情報: 横方向への展開

獲得した情報を構成する要素を変数として、類似情報を比較的に獲得する。例えば、技術調査レポートを作成するユーザが、新聞記事検索を行なって、「A 社が低価格の X 製品を開発した」という情報を検索した場合、詳細情報は新聞記事本文で得ることができるが、「他社はどうなっているか」、「A 社の Y 製品はどうなっているか」などの情報を 2 次的検索として行なう。いわば、1 次情報からの水平方向、横方向への展開がある。

鳥瞰情報: 情報全体像の把握

獲得した情報が数量的に膨大な場合、全体像を鳥瞰的に把握する。例えば、製品開発情報の集合が得られた場合、いつ頃、どのような種類の組織が、どのような分野に関して、どのくらいの件数を発表しているのかといった集合全体を把握するという展開がある。

これらの情報展開に関しては、現状の情報検索技術は必ずしも、有効な手段を提供していない。例えば、情報

検索の結果、ある注目すべき事件があって、そこに至るまでの事件の流れを知りたいと思っても、関連する一連のニュースだけを拾い読みすることは簡単ではない。また、製品発表のニュースなどから、他社の類似製品の動向についても知りたいと思っても、それを探すには、改めて検索を行わなければならない。さらに、配信される情報の量が多い場合には、全体の傾向がつかめず、目的の情報を探しにくくなるという問題もある。

そもそも、複数のキーワードの組み合わせで情報検索しようとしても、それらが異なる部分に現れる文書もヒットしてしまいノイズとなる情報も多い。例えば、**NEC& 半導体 & 生産** というキーワードで新聞記事を検索すると、「NEC が～と技術提携し、～が半導体を生産する。」というように 3 つのキーワードの間に直接的なつながりが存在しない文も、検索されることになる。このように、キーワードの存在をベースとした情報検索は、上述した 3 方向への情報展開を図る場合にはノイズを発生する。

3 5W1H 情報による分類・ナビゲーション

キーワードの組み合わせによるノイズの問題を解決するためには、キーワードの文中での役割に着目し、5W1H 情報を用いることが有効である。新聞記事 30 万文に対して、2 つのキーワードを AND 条件で検索した場合と、A と B の間に構文的な係り受けの関係がある場合で検索した場合を比較すると、約半分ほどに情報が絞り込まれることが報告されている [5]。そこで、5W1H 情報を利用して、ユーザーが読みたい情報へとユーザーをナビゲートする手法について説明する (図 1)。

5W1H 分類・ナビゲーションは、エピソード抽出、多視点分類、情報鳥瞰の 3 つ機能によって、情報の縦方向への展開、横方向への展開、全体像の把握を可能とする。これらの機能を用いることによって、1) ある出来事に至るまでの経緯が簡単に分からない 2) ある出来事に関連する他の情報を簡単に探すことができない 3) 多量の情報の中から必要な情報に素早くアクセスできない、という 3 つの問題の解決を図る。

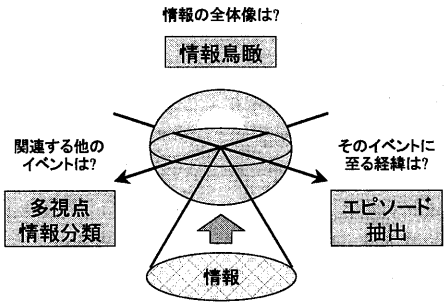

図 1: 5W1H 分類・ナビゲーションモデル

3.1 エピソード抽出

5W1H の条件を指定して検索を行うことで、ある出来事について述べている文書だけを抽出し、結果を時間順に並べて提示する。この結果、その出来事に関するこれまでの経緯をエピソード的に読むことができるようになる。例えば、Who 要素に NEC、What 要素に PDP、How 要素に開発を含む文書を検索し、時間順に並べることで、NEC の PDP の開発に関する出来事をエピソードとして抽出することができる。単純なキーワードによる条件指定では、NEC と PDP と開発との関係を指定できず、それらが異なる文脈に現れる文書もヒットしてしまうが、5W1H の条件指定でそれを防ぐことができる。

3.2 多視点分類

ある情報に関連する情報を、各 5W1H のキーワードごとに、そのキーワードを含むかどうかで分類する。この結果、5W1H による機能的な観点から関連情報にアクセスできる。5W1H の各要素についての分類を組み合わせると、6 次元空間上に分類することになるが、ユーザーには、基準となる要素と別な要素の組み合わせによる 2 次元の分類結果を提示し、基準となる軸を変えた分類結果を次々と切り替えて表示できるようにする。これにより、ユーザーは、わかりやすい 2 次元のマトリクス形式で分類を見ることができ、視点を切り替えていくことで、連想的に関連情報を探すことができる。

3.3 情報鳥瞰

5W1H のそれぞれの軸上に要素が多数存在して、多視点分類結果が膨大なマトリクスとなる場合、シソーラスを用いることによって、5W1H 要素をそれらの上位概念で代表させて分類する。キーワードごとの細かな分類ではなく、概念ごとの大まかな分類を生成する。指定された部分については、細かい分類を生成することもできる。この機能によって、多くの情報を分類対象とする場合でも、ユーザーは、適度なレベルで分類結果を参照し、情報全体を鳥瞰しながら、必要な情報を探すことができる。

4 5W1H 要素の抽出

5W1H 要素の抽出は、形態素解析と辞書情報を基にした浅い解析手法、表層格指向パーシングによって行なう (図 2)。表層格指向パーシングによって、頑健で効率的な 5W1H 抽出が可能である。

5W1H 抽出は、固有名詞抽出、時間表現抽出、格要素抽出の 3 つのステップで行なう。各文に対して形態素解析および表層格の解析を行った後、6 万語の固有名詞辞書を用いて、人名・組織名を Who 要素として、地名を Where 要素として抽出する。例えば、「NEC は中国で半導体を生産」という文からは、NEC と中国が Who 要素と Where 要素として抽出される。When・Where・Why 要素については、「～年～月に」・「～地区に」・「～のため」等の、日時や場所、理由を表す特徴的な表現に着目して抽出する。例えば、「97 年 4 月に新製品を発表」という文では、97 年 4 月が When 要素として抽出される。基本的には、約 1 万 5 千語の動詞辞書を用いて、動詞と助詞のパターンに着目して、文中の動詞を How 要素に、が格要素を Who 要素に、を格要素を What 要素とする。例えば、「A 社が VOD システムを受注」という文では、A 社、VOD システム、受注が、それぞれ、Who 要素、What 要素、How 要素として抽出される。

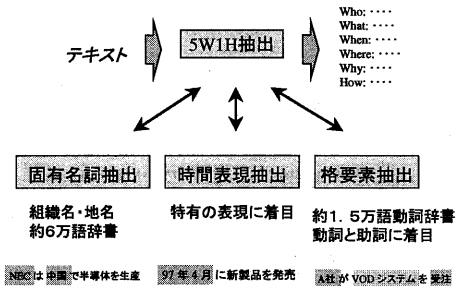

図 2: 表層格指向パーシング

5 5W1H 分類・ナビゲーションによる情報活用プラットフォーム

新聞記事 8000 件を対象として、5W1H に基づいて分類・ナビゲーションを行う情報活用プラットフォームを構築した (図 3)。ここでは、指定された情報源から情報収集ロボットによって情報が収集され情報 DB に毎日格納されていく。5W1H 情報抽出モジュールは、収集した情報から 5W1H 情報を抽出し 5W1H インデクスを生成する。ここでは、新聞記事ヘッドラインから Who・What・How の 3 種類について 5W1H 要素を抽出している。5W1H インデクス情報を基に、情報分類モジュール、エピソード抽出モジュール、鳥瞰情報生成モジュールが、WWW サーバーを介してユーザに分類・ナビゲーション機能を提供する。フィルタリングサービス部は、ユーザプロファイルを参照して情報収集ロボットによって集められた情報からフィルタリングしてユーザに配信する。ユーザは配信された情報から、5W1H 分類・ナビゲーションによって自分の読み進みたい方向へと情報を獲得していくことができる。

5.1 新聞記事に対する適用

エピソード抽出では、指定されたヘッドラインから Who・What・How 要素を抽出し、それと同じ Who・What・How 要素の組を持つ文書を関連エピソードとして検索して、結果を時間順に並べて表示する。図 4 に、「NEC 半導体部門の生産予測を 18% 増と発表」という

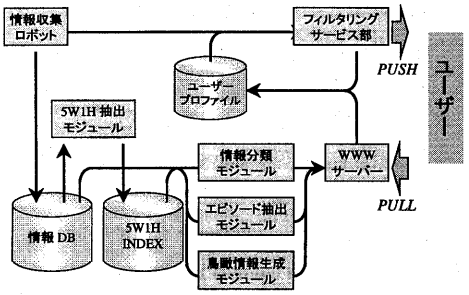

図 3: 情報活用プラットフォーム

ヘッドラインから抽出した5W1H要素のうち、NEC・半導体・生産というWho・What・How要素の組に対してエピソード抽出を行った結果を示す。抽出された記事の見出しを読むことで、NECの半導体の生産が18%増に至るまでの経緯を知ることができる。

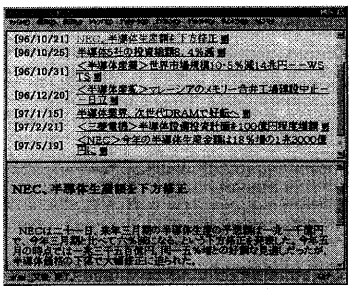

図 4: 5W1H によるエピソード抽出

多視点分類では、抽出した3種類の5W1H要素のうち、2つに共通のキーワードを含むものを関連ヘッドラインとして検索し、検索結果をキーワードごとに分類する。図5と図6に、「NEC・A社・B社 暗号化データの回復技術を開発へ」というヘッドラインから多視点分類を行った結果を示す。

図5は、Whoの軸を基準とした場合の結果である。分類軸をWhatに切り替えると図6となる。図5では、例えば、NEC・A社・B社といったWhoの視点から、各会社の暗号化に関する情報件数を知り、それらの情報に容易にアクセスできる。図5の画面で、Whatを

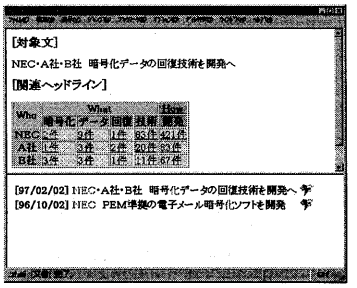

図 5: Whoの視点からの分類

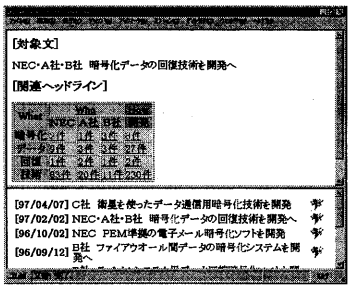

図 6: Whatの視点からの分類

クリックすると、Whatの視点から分類することができる(図6)。Whatの視点から、例えば、暗号化の開発については8件のデータがあり、それらの情報にも容易にアクセスすることができる。

情報鳥瞰では、Who 要素に出現する約2800の企業を業種別に分類したシソーラスと、What要素に出現する約1000のキーワードを技術分野別に分類したシソーラスを利用して、Who・What 要素の各キーワードを統合し、鳥瞰的な分類を生成する。How 要素については、キーワードの種類が少ないことから、高頻度で出現するキーワード8語で分類している。図7に約400件のヘッドラインに対する情報鳥瞰の結果を示す。Who・What 要素のキーワードを階層的なシソーラス構造として扱うことで、最初は荒い分類を提示し、必要な部分だけ展開して細かい分類を見せることができる(図8)。

5.2 セールスレポートに対する適用

実際のセールスレポートから実験用に作成した約2500件のセールスレポートに対して5W1H分類・ナビゲー

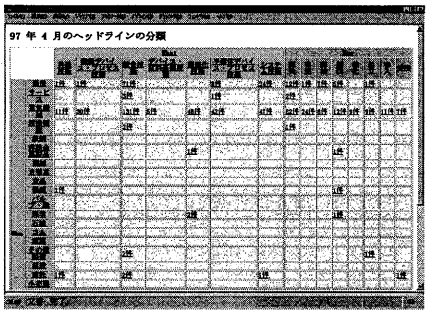

図 7: シソーラスを利用した情報鳥瞰

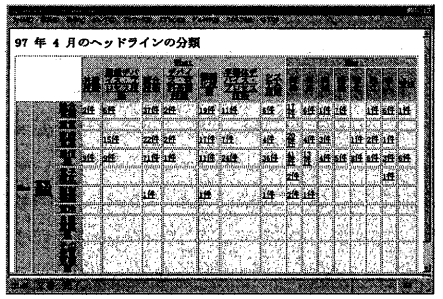

図 8: シソーラスを利用した情報鳥瞰: 細分類

ションを適用した例を示す。セールスレポートは、営業マンが担当するユーザの要望などを報告したものである。このレポートは、メッセージアベニューという配信サービスとして関係者に配信される。セールスレポートは、5W1H 情報抽出によって、日時、ユーザー名、対応支店名、機種名、用件が抽出される。新聞記事データと同様に、それぞれの視点を切り替えることによって多視点情報分類を実現し、ユーザーをナビゲートする。

図 9は、一カ月分 (97 年 4 月分) のユーザーレポートの一覧である。画面上部のフレームにユーザーレポートのリストが、画面下部のフレームに選択されたユーザーレポートの内容が表示される。図 9の例では、リストの 3 番目のレポートを表示している。このレポートにおいて、○×建築がユーザー名、若葉台支店が対応支店名、XX282 ソフトが機種名、2000 年問題が用件に対応する。この画面から、例えば、○×建築殿という部分をクリックすることでユーザー○×建築に関するレポートの多視点分類を行うことができる。図 10上画面が、ユー

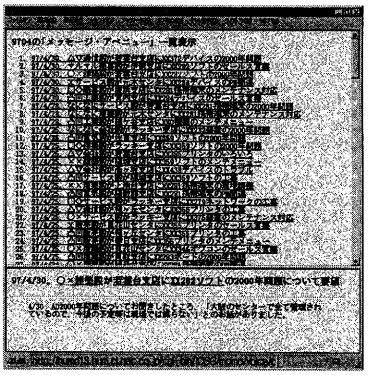

図 9: セールスレポートの一覧

ザー○×建築に関するレポートの分類である。これにより、○×建築については、若葉台支店以外にも、上作延支店が XX554 ネットワークに関して対応していることが分かる。実際にどのような対応を行っているかについては、上作延支店対 XX554 ネットワークの部分のセルをクリックすることで、対応するユーザーレポートのリストを得ることができる (図 10下画面)。一方、上作延支店という部分をクリックすることで、今度は、視点を

図 10: ユーザ先からの分類

上作延支店に移し、上作延支店に関するレポートの分類を得ることもできる (図 11)。これにより、上作延支店が他にどんなユーザーを持っているのか、あるいは、他にどの機種を扱っているかを知ることができる。

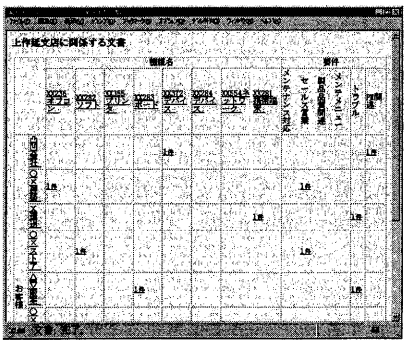

図 11: 支店からの分類

6 おわりに

本稿では、5W1H 情報によって、エピソード抽出・多視点分類・情報鳥瞰を機能とする情報分類・ナビゲーションの手法を提案した。これらの機能により、1) ある出来事に至るまでの経緯が簡単に分からない 2) ある出来事に関連する他の情報を簡単に探すことができない 3) 多量の情報の中から必要な情報に素早くアクセスできない、という課題の解決を図った。新聞記事 8000 件とセールスレポート 2500 件を対象として、情報活用プラットフォームを構築し、5W1H 分類・ナビゲーションが、上述の課題の解決に有効であることを確認した。今後、以下の改良を行ないシステムとしての完成度を高めていく。

- 5W1H 抽出: 5W1H 抽出は、ヘッドラインやタイトルに関して 9 割の精度で抽出可能であるが、一般文では精度は半分程度である。埋め込み文や括弧の処理などよりきめ細かい処理を行ない、表層格指向パーシングの改良を図る。
- エピソード抽出: 5W1H 要素の中でエピソードとして抽出する条件をユーザが指定し調節可能とする。
- 多視点分類: 分類項目をユーザが指定可能とし、また、複合的な条件からの分類も行なう。
- 情報活用プラットフォーム: 複合オントロジを利用した情報集配信サービス MIIDAS との統合により、複合オントロジからのユーザープロファイル情報の獲得とナビゲーション履歴の利用による個人適合化を図る。

参考文献

[1] 坂本仁, “情報利用に向けた自然言語処理技術,” JEIDA 自然言語処理技術に関する講演会,1997/6/17

[2] 奥村明俊, 他, “オントロジによる多次元情報集配信,” 人工知能学会第 11 回全国大会, pp.368–339,1997

[3] 奥村明俊, 池田崇博, 村木一至, “MIIDAS: 情報の選別的共有のためのオントロジ構築とその増進的学習,” 情報処理学会第 55 回全国大会, 5Q-08 (1997).

[4] 池田崇博, 奥村明俊, 村木一至, “5W1H 情報を利用する情報分類・ナビゲーション,” 人工知能学会第 11 回全国大会, pp.370–371 (1997).

[5] Kenji Satoh, Kazunori Muraki, “Penstation for Idea Processing,” Natural Language Processing Pacific Rim Symposium (NLPRS'93),December 1993