MIIDAS: 情報の選別と Easy Reading のためのエピソード

5Q－10

池田 崇博　奥村 明俊　村木 一至
NEC C&C メディア研究所

1　はじめに

インターネットの急速な普及とともに、多くの情報を簡単に手にすることができるようになったが、その情報は、内容・質ともに千差万別であり、有用な情報だけを取り出すことは難しい。そこで、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ーを利用して、情報を選別し、配信するサービス MIIDAS (Multi-Indexing Information Dissemination & Acquisition Service) の実現を目指している [1]。

情報集配信サービスにより、ユーザーには、欲しい情報だけが選択的に配信されるようになる。しかしながら、配信結果は断片的な情報の集合であるため、ユーザーが関連するエピソードを内容に従って読み進めていくことは困難である。例えば、ニュースの配信を受けている場合、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。複数のキーワードの組み合わせで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う。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、改めて煩雑な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。

MIIDAS では、1) 配信された情報に対して、それに至るまでの経緯が簡単に分からない 2) 配信された情報から、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 つの問題を解決するために、文章中のエピソードの内容を表す Who(だれが)・When(いつ)・Where(どこで)・What(なにを)・Why(なぜ)・How (どうした) の 5W1H に着目して情報を分類・整理する機能を提供する [2]。

2　テキストからの 5W1H 要素の抽出

5W1H の観点から情報を分類・整理するために、テキストから 5W1H 要素に分けてキーワードを抽出しておく。頑健で効率的な 5W1H 抽出を実現するために、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、動詞と助詞のパターンに着目して 5W1H 情報を抽出する。基本的に、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 要素にしている。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。さらに、固有名詞に着目し、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ことで、格パターンの解析を補っている。

MIIDAS: Information selection by episodic easy-reading
Takahiro Ikeda, Akitoshi Okumura, and Kazunori Muraki
NEC C&C Media Research Labolatories

3　5W1H による情報分類・ナビゲーション

抽出した 5W1H 要素を基に、以下の 3 つの情報分類・整理機能を提供し、ユーザーをナビゲートする。

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。

多視点分類　ある情報に関連するさまざまなエピソードを、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文書にアクセスできるようになる。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 次元の表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、連想的に関連文書を探していくことも可能になる。

情報鳥瞰　シソーラスを参照して、複数の 5W1H 要素をそれらの上位概念で代表させて扱うことで、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ようにする。これにより、分類対象となる文書が多い場合でも、ユーザーは、適度なレベルの分類を見ることができ、対象文書全体を鳥瞰しながら、目的の文書を探すことができるようになる。

4　分類・ナビゲーション機構の試作

実際に、新聞記事を 30 字程度に要約した新聞記事ヘッドラインを対象として、5W1H に基づいて分類・ナビゲーションを行う機構を試作した。ここでは、

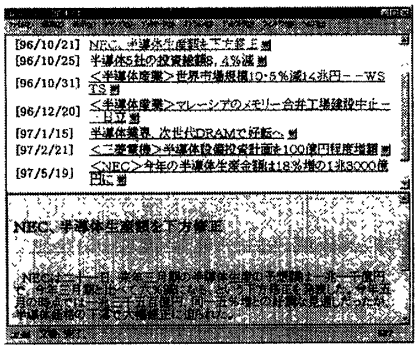

図 1: 5W1H によるエピソード抽出

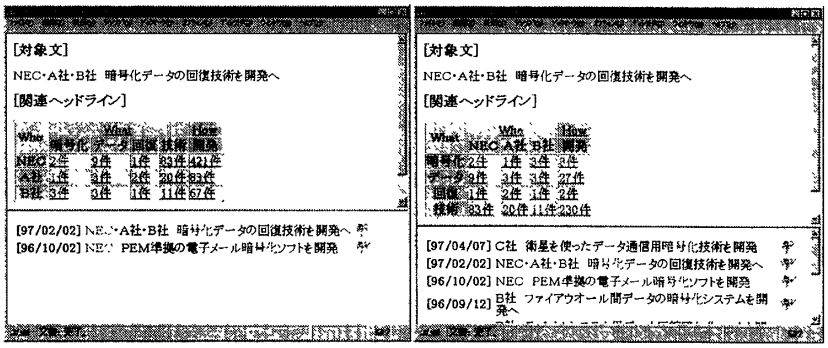

図 2: 多視点分類における 5W1H 視点の切り替え

Who・What・How の 3 種類について 5W1H 要素を抽出し、前節で述べた 3 つの機能を実現した。

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 1に、「NEC 半導体部門の生産予測を 18% 増と発表」という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。

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 2 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。左の画面は、Who の軸を基準とした場合の結果である。分類軸を What に切り替えると右の画面になる。左の画面を見ることで、例えば、NEC・A 社・B 社といった Who の視点から、各会社の暗号化についての過去のヘッドラインに容易にアクセスできる。一方、右の画面を見ることで、例えば、暗号化の開発についての過去のヘッドライン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 語だけで分類している。図 3 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ようになっている。

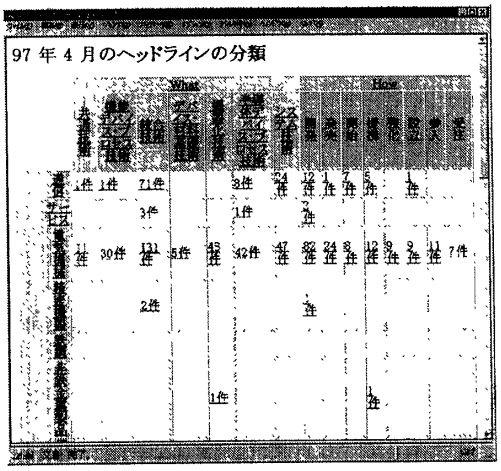

図 3: シソーラスを利用した情報鳥瞰

5　まとめ

本稿では、配信された情報から、それに至るまでのエピソードや、他の関連する情報を簡単に手にすることができ、また、多量の情報の中から、必要な情報に素早くアクセスできるようにユーザーをナビゲートする機能として、5W1H によるエピソード抽出・多視点分類・情報鳥瞰を行うことを提案した。また、新聞記事ヘッドラインを対象とする分類・ナビゲーション機構を試作し、これらの機能により、当初の 3 つの課題を解決できることを確認した。今後、対象を一般の文書に広げていくとともに、エピソード抽出条件の調節や、分類項目への複合条件の設定等の機能を組み込み、分類・ナビゲーションシステムとしての完成度を高めていく予定である。また、複合オントロジーを利用した情報集配信サービス MIIDAS への組み込みをはかり、複合オントロジーからのユーザー情報の獲得とナビゲーションへの利用についても検討していく予定である。

参考文献

[1] 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).

[2] 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).