エピソード検索における適合レベルの調整

池田 崇博　佐藤 研治　奥村 明俊*
NEC C&C メディア研究所†
e-mail: {t-ikeda, satoh, okumura}@ccm.cl.nec.co.jp

1 はじめに

インターネットの急速な普及や大容量のストレージの普及とともに、膨大な量の情報を手にする機会が多くなっている。近年、このような大規模なテキストに対して、高速に検索を行う技術が確立され、キーワードベースの検索システムや、ユーザのプロファイルに適合する情報配信サービスなどにより、自分に必要な情報を検索したり選別したりできるようになった。しかしながら、一般に、情報の検索や選別自身は、ユーザの目的でないにも関わらず、そのために多くの時間と手間が費やされているのが現状である。

例えば、ニュース記事の配信を受けている場合に、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うと、配信された記事中のキーワードを抜き出して、改めて検索を行わなければならない。しかも、単純にキーワードの and、or の条件による検索を行っただけでは、関連する一連のニュースだけを拾い読みすることができないのが普通である。例えば、ある会社 A 社の半導体の生産についてのこれまでの経緯を知りたいという場合に「A 社 & 半導体 & 生産」という条件で検索を行うと、「A 社が生産するコンピュータは、B 社の半導体を使っている。」という文でもヒットしてしまう。結果として、A 社の半導体の生産に関する文書以外のノイズが乗ることになる。

このような問題を解決するために、これまでに、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; 本稿では、述語要素を How として扱う) の 5W1H に着目し、検索時に 5W1H の条件を考慮することで検索ノイズを減らす手法 (エピソード検索) を提案してきた [1, 2]。これは、検索式としてキーワードだけではなく、キーワードの 5W1H としての役割を指定するようにし、検索式中のキーワードが、指定されたものと同じ 5W1H の役割で出現する文書だけを検索結果とするものである。エピソード検索は、5W1H の条件として特定の出来事の 5W1H 内容を指定することで、その出来事に関するこれまでのエピソードを抽出することができるため、特定の出来事に対する時間的な経緯を知る手段として有効に活用できることが分かっている。

本エピソード検索技術を用いると、入力した条件に正確に一致するエピソードが抽出されるが、ユーザが、検索意図を正確に反映した検索条件をいつも入力できるとは限らない。例えば、文を入力として、その文中の 5W1H を基にエピソード検索を行う場合がある。このとき、ユーザの検索意図が、1 つの文中に表れる複数の 5W1H のうちの一部だけであると、検索条件としての文は、検索意図と離れたものになってしまう。このような課題に対応するためには、入力した文と検索対象の文との間での 5W1H の一致度を、ユーザがコントロールできるようにする必要がある。

そこで、本稿では、ある出来事に関するエピソードとして、その出来事の内容と 5W1H が完全に一致するもの以外に、5W1H の一部が一致しているものも検索しておき、ユーザが自由に 5W1H の一致度 (適合レベル) を選択できるようにエピソード検索を拡張する。検索式と対象文書との 5W1H の適合レベルをいくつか設定しておき、予め各レベルごとに結果の件数を求めて提示することで、ユーザが再検索を行うことなく、検索意図や結果の量に合った適合レベルの結果を得られるようにする。

以下では、まず 5W1H を利用したエピソード検索について説明し、続いてエピソード検索における 5W1H の適合レベルの調整機能の詳細について述べる。

2 5W1H を利用したエピソード検索

2.1 5W1H 要素の抽出

5W1H 要素の抽出は、CBSP (Case-Based Shallow Parsing) モデルに従って行う。CBSP モデルは、形態素解析を行い各単語に品詞情報を付与したテキストに対し、語彙情報・字句のパターン・助詞の情報を用いて 5W1H 解析を行うモデルで、浅い解析により、頑健で効率的な解析を実現している。基本的には以下の 3 ステップより構成される。

(1) 固有名詞の抽出
固有名詞のうち、人名・組織名を Who 要素として、

*Takahiro Ikeda, Kenji Satoh, and Akitoshi Okumura
†NEC C&C Media Research Labolatories

表 1: Who, What, How の各要素および全体での抽出結果の評価

	Who 要素			What 要素			How 要素			
	存在	非存在	計	存在	非存在	計	存在	非存在	計	全体
正解	5423	71	5494	5653	50	5703	6042	5	6047	5270
誤り	414	490	904	681	14	695	55	296	351	1128
合計	5837	561	6398	6334	64	6398	6097	301	6398	6396
精度	92.9%	12.7%	85.9%	89.2%	78.1%	89.1%	99.1%	1.7%	94.5%	82.4%

地名は Where 要素として抽出する。これにより、例えば、「NEC が中国で半導体を生産する。」という文からは、Who 要素として NEC が、Where 要素として中国が抽出される。現在、約 6 万語の固有名詞辞書を利用している。

(2) 特徴的表現のパターンマッチ

特徴的なパターンに着目して、人名・組織名（Who 要素）、日時（When 要素）を抽出する。例えば、「株式会社××」、「××大学」などのように「株式会社」が頭に付く語や大学が後に続く語は、会社名や大学名と考えられるため、これらを Who 要素として抽出する。また、「平成××年×月」、「××/××/××」のようなパターンに当てはまる語は日時を示していると考えられるため、これらは When 要素として抽出する。現在、人名・組織名のパターンを約 100 種類、日時のパターンを約 20 種類用意して、解析に利用している。

(3) 表層格解析

上記（1）、（2）のステップで抽出されなかった名詞は、その名詞に続く助詞等の情報を基に、どの 5W1H 要素に対応するのかを決定する。例えば、「が」および「は」が後に続く語は Who 要素とし、「を」および「に」が後に続く語は What 要素とする。動詞は How 要素として抽出する。

CBSP による解析で、約 6,400 件の新聞記事ヘッドラインから、実際に Who、What、How 要素を抽出した結果について分析した結果を表 1 に示す。この表では、新聞記事ヘッドラインにおいて、Who、What、How の各要素が実際に存在している場合と存在していない場合のそれぞれについて、それらの要素が正しく抽出された文の数と正しく抽出されなかった文の数をまとめている。これによると、Who、What、How の各要素が実際に存在している場合には、ほぼ 90%以上の文から各要素が正しく抽出できている。しかしながら、各要素が実際に存在していない場合には、高い精度が得られていない。これは、要素が実際に存在していない場合でも、別な語をその要素として抽出してしまう傾向があるためである。これにより、関係のない語が 5W1H 要素として抽出されることになるが、正しい 5W1H 要素が落ちるわけではないので、5W1H を利用した検索において適合率を下げることはなく、実際には、大きな問題とはならない。要素がある場合とない場合との平均では、85%から 95%の精度となっており、全体でも 82.4%の精度が得られている。

2.2 エピソード検索

ある出来事の 5W1H の内容を検索条件として指定し、それと同一の 5W1H の内容を含む文書を検索することで、その出来事について述べている文書だけを抽出することができる。抽出した結果を時間（When）順に並べると、その出来事に関するこれまでの経緯が時間的な流れに沿って表示されるため、その出来事にまつわるエピソードとして順に読むことができる。そこで、本稿では、このような検索方式をエピソード検索と呼んでいる。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。

新聞記事ヘッドラインの配信サービスなどを受けているとき、その配信された内容について、より詳しい情報、例えば、その記事で述べられている出来事に至るまでの経緯を知りたいと思うことがある。このような場合に、ヘッドライン中で述べられている 5W1H を基に過去の新聞記事に対してエピソード検索を行うことで、これまでの流れを知ることができる。図 1 に「NEC が半導体の生産を予定より 18%増」という新聞記事ヘッドラインに含まれる、Who 要素が「NEC」、What 要素が「半導体」、How 要素が「生産」という 5W1H 関係を基に、NEC の

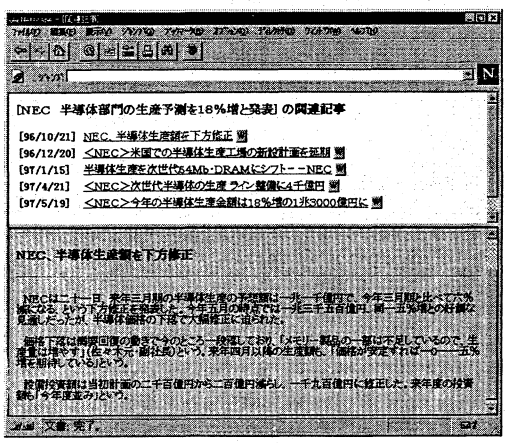

図 1: 5W1H によるエピソード抽出

半導体の生産に関するエピソードを検索して表示した例を示す。画面上部のフレームにヒットした記事の見出しのリストが表示されているが、これによると、96 年後半には不況だった半導体市場が、97 年初頭に DRAM の世代交代があり、その後好況に転じたことがエピソードとして読みとれる。

この例のように、ある文を基にエピソード検索を行う場合、入力された文から 5W1H 要素を抽出して、それと同一の 5W1H の内容を含む文書を検索することで、ユーザが改めて 5W1H の条件を指定することなく、基にする文を指定するだけでエピソード検索を行うようにすることもできる。

3 エピソード検索における適合レベルの調整

前章では、もっとも基本的な場合として、入力された 5W1H と同一の 5W1H を含む文書からなるエピソードについてエピソード検索を定義したが、必ずしもすべての 5W1H を共有していなければ、エピソードとして意味がないというわけではない。例えば、NEC の半導体の生産に関するエピソードを知りたいという場合でも、「NEC が半導体を生産する」という内容に関することだけが知りたい場合もあれば、「NEC が半導体を開発する」あるいは「NEC が半導体を出荷する」という内容に関することも知りたいという場合もある。特に、How 要素の場合、異なる単語でも類似する意味を表すことが多く、How 要素が一致していない文書にも関連する内容が書かれていることが多い。

また、1 つの文には、一般に 5W1H が複数セット含まれるため、特に文を基にエピソード検索を行う場合、入力と対象文書との間での 5W1H の一致性が低くてもよい場合がある。例えば「A 社が製品 X を発売、B 社は製品 Y を発売。」という文には、Who 要素が「A 社」、What 要素が「製品 X」、How 要素が「発売」という 5W1H と、Who 要素が「B 社」、What 要素が「製品 Y」、How 要素が「発売」という 5W1H が含まれているが、この文に関するエピソードとして、「A 社が製品 X を発売」ということに関するものだけが欲しいことや、「B 社が製品 Y を発売」ということに関するものだけが欲しいことがある。しかしながら、「A 社が製品 X を搭載する PC を発売。」という文には、Who 要素が「A 社」、What 要素が「PC」、How 要素が「発売」という 5W1H と、Who 要素が「PC」What 要素が「製品 X」、How 要素が「搭載」という 5W1H が含まれているが、この場合に、「A 社が PC を発売」という 5W1H や「PC が製品 X を搭載」という 5W1H だけでエピソードを検索すると、元の文の内容とは関連の薄いものが多く混入してしまうおそれがある。

そこで、与えられた文と一部の 5W1H 要素を共有する文を含む文書の集合として、エピソードの概念を一般化し、文と文との 5W1H の適合レベルを複数段階用意して、適合レベルに応じてエピソードを検索するようにエピソード検索を拡張する。その上で、適合レベルごとに検索を予め実行しておき、ユーザがその結果を見て、検索条件が適当なものや検索結果の量が適当なものなど、検索意図に合う結果を再検索することなく得られるようにする。

入力文と検索対象の文との間での 5W1H の適合レベルとしては、以下の 5 段階を定義する。

(1) 入力文中に含まれる 5W1H すべてが対象文中に含まれる
(2) 入力文中に含まれる 5W1H の少なくとも 1 つが対象文中に含まれる
(3) 入力文中に含まれる各 5W1H から How 要素を覗いたものが少なくとも 1 つ対象文中に含まれる
(4) 入力文中に含まれる各 5W1H のうちの 2 要素を組み合わせたものが少なくとも 1 つ対象文中に含まれる
(5) 入力文中に含まれる各 5W1H 要素が少なくとも 1 つ対象文中に含まれる

レベル 1 は、入力文中の 5W1H すべてについて対象文中に書かれているという条件に、レベル 2 は、入力文中

の 5W1H のうちどれか 1 つについては対象文中で書かれているという条件になっている。レベル 3 は、入力文中の 5W1H で How 要素にあたるものが、対象文と一致していなくても適合したとみなすものである。How 要素は、異なる単語でも同じ内容を表していることが多く、通常の検索要求では、Who や What さえ一致していれば十分なことが多いので、How を外したものを 1 つのレベルとした。レベル 4 は、入力文中で 1 つの 5W1H として内容を表現している各要素のうち、対象文と 2 要素が一致していれば、適合するとみなす条件で、レベル 5 は、1 要素の一致でも適合するとみなす条件である。

これらのレベルは、上位のレベルで適合する文書は下位のレベルでも適合するように定義している。このため下位のレベルでの結果ほど、より多くの広範な内容の文書を含むことになる。予め各レベルごとに検索を行っておき、各レベルでの適合文書数をユーザに提示することにより、ユーザは、自分の検索目的や結果の量に合わせて適切なレベルを選択することで、再検索することなく目的の結果を得ることができる。

例えば、Who 要素が X で、What 要素が Y で、How 要素が Z の 5W1H を含むという条件を (Who:X, What:Y, How:Z) と表すことにすると、「A が B を C して、P が Q を R した」という文に対して、それぞれの適合レベルに対する検索条件は、

(1) (Who:X, What:Y, How:Z)
    and (Who:P, What:Q, How:R)
(2) (Who:X, What:Y, How:Z)
    or (Who:P, What:Q, How:R)
(3) (Who:X, What:Y) or (Who:P, What:Q)
(4) (Who:X, What:Y) or (Who:X, How:Z)
    or (What:Y, How:Z) or (Who:P, What:Q)
    or (Who:P, How:R) or (What:Q, How:R)
(5) (Who:X) or (What:Y) or (How:Z) or (Who:P)
    or (What:Q) or (How:R)

となる。レベル 4 や 5 では、適合文書数が非常に多くなると考えられるが、適合文書を、適合した 5W1H 要素の数等でランキングすることで、ユーザが適合度の高いものから順に結果を見られるようにする。

2 章で用いた約 6,400 件の新聞記事ヘッドラインを対象として、「NEC が PDP を開発。」のような入力文 10 種類に対して、各レベルごとの適合文書数を求め、平均した結果を表 2 に示す。新聞記事ヘッドライン中には、5W1H が 1 セットしか出現しないので、入力文も 5W1H が 1 セットからなるものとした。このため、レベル 1 とレベル 2 の適合文書数は等しくなっている。

表 2: 各適合レベルごとの適合文書数の平均

レベル	1	2	3	4	5
適合文書数	2.3	2.3	10.7	109.0	1447.7

表 2 によれば、レベルが大きくなるにつれて順々に結果の文書の数が大きくなっていく様子が確認できる。各レベルごとの実際の検索条件とその検索結果による適合文書数を予めユーザに提示することにより、ユーザは、レベル 2 の条件が目的に近いからレベル 2 の結果を見よう、あるいは、レベル 2 の 2 件では情報が少なすぎるのでレベル 3 の 10 件を見ようというように、検索条件や適合文書数に応じてレベルを選択できる。

4 まとめ

本稿では、文章中のだれが、いつ、どこで、なにを、なぜ、どうしたという 5W1H の要素に着目し、検索時に 5W1H の適合性を考慮するエピソード検索を拡張する手法について述べた。5W1H によって出来事の内容を表すことができるため、エピソード検索で特定の 5W1H を含む文書を時間順に並べることによって、その出来事に関する時間的な経緯をエピソードとして読むことができる。しかし、エピソードとしては、5W1H のすべてが完全に一致する文書以外にも有効なものも存在し、必ずしも 5W1H の一致度が高い文書だけが有効とは限らない。そこで、入力文と検索対象の文との間の 5W1H の適合レベルを 5 段階定義し、それぞれのレベルで適合する文書数を予め検索してユーザに提示することで、ユーザが適合文書数等に応じて適合レベルを選択し、再検索を行うことなく適切な結果を得られる方法を提案した。今後、現状のエピソード検索システムに適合レベルの調整機能を実装し、評価を行っていく。また、各適合レベルでの適合文書の適切なランキングについても併せて検討する。

参考文献

[1] 池田崇博, 奥村明俊, 村木一至, "MIIDAS: 情報の選別と Easy Reading のためのエピソード," 情報処理学会第 55 回全国大会, 3, pp.244–245 (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).