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

2006－NL－172（1）
2006／3／27

時系列情報を利用した複合語キーワードの抽出

村上 明子　渡辺 日出雄
日本 IBM 東京基礎研究所
〒 242-8502 大和市下鶴間 1623-14
akikom@jp.ibm.com, hiwat@jp.ibm.com

テキストマイニングなどのテキスト分析では、名詞などをキーワードとして分析の単位としている。我々は分析の一つの概念と見なせるこの単位を Mining Concept と呼んでいる。Mining Concept はその定義によってテキスト分析の結果に大きく影響する。我々は今回、Mining Concept の一つの種類として時系列発生的 Mining Concept(FOMIC) を定義した。FOMIC とは、文書内で名詞の複合語を形成している単語群で、各々の単語頻度の時系列推移が同期しているものである。この Mining Concept は、話題の時間推移に同期しているといえる。本論文では、我々の定義した FOMIC の抽出方法を示し、その有効性を議論する。

キーワード抽出　時系列解析　複合名詞　テキストマイニング

Mining Concept: Compound Noun Recognition referring to Chronological Fluctuation of Occurrence

Akiko Murakami　Hideo Watanabe
Tokyo Research Laboratory, IBM Japan
1623-14 Shimotsuruma Yamatoshi
Kanagawaken 242-8502 Japan
akikom@jp.ibm.com, hiwat@jp.ibm.com

Abstract

In text mining, we want to use not only single nouns but also compound nouns and noun phrases as an unit of analysis. We usually call this unit of analysis a “mining concept.” We define a subclass of mining concepts called *Fluctuation of Occurrence based MIning Concept (FOMIC)*. Our approach is based on the idea that a compound noun whose constituents have the similar temporal frequency distribution can be considered as a FOMIC. In this paper, we propose a method to extract FOMICs and discuss the effectiveness of FOMIC as a mining concept.

Key Words　Keyword Extraction, Chronological Analysis, Compound Noun, Text Mining

1　はじめに

昨今、インターネットにおける掲示板やブログなどの出現により、消費者が企業や商品などの感想や苦情などをコールセンターなどのチャネルを通さずに表現することが可能となった。このようなテキストはコールセンターに寄せられた意見やアンケートの結果よりも手軽に、大量に入手することが可能であり、今後これらの分析の必要性は大きくなってくる。

これらを対象とした分析の一つに、テキストマイニング (Nasukawa, 2001) があげられる。これは、単語を一つの分析単位とみなし、その分析単位の出現頻度の時間的変化や、共起関係などを見て分析する手法である。テキストマイニングにおいては、この分析単位によって結果が大きく変わってくるため、この分析単位を決めることは非常に重要である。我々はこの分析単位を Mining Concept と呼んでいる。

多くのツールはこの分析の単位を形態素解析に委ねている。しかし、テキスト分析を行う際には形態素解析の結果である単語（以後複合語と区別するため単単語と呼ぶ）だけではなく、その単単語から構成される複合語やフレーズといった拡張した概念も一つの単位として扱った方がより正確な分析が可能になる。

例として「インフルエンザ」と「鳥インフルエンザ」という二つの Mining Concept を考える。図 1 に実際の 2003 年に書かれたインターネットの文書における「インフルエンザ」と「鳥インフルエンザ」の文字列の出現回数を示す[1]。

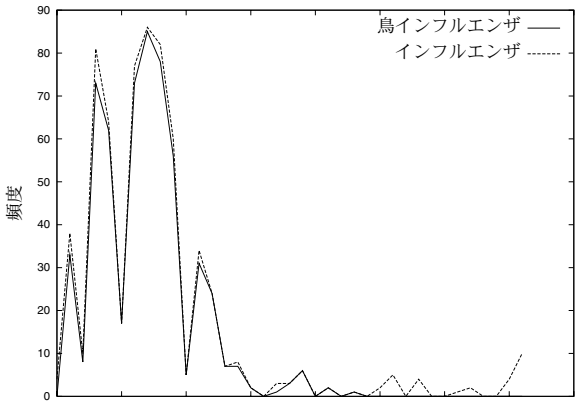

図 1: 「インフルエンザ」、「鳥インフルエンザ」の頻度の時間推移

図 1 を見ると、この文書の中で出てくる「インフルエンザ」の大部分がいつの時点でも「鳥インフルエンザ」の部分文字列として出現していることがわかる。しかし、通常「鳥インフルエンザ」は鳥から鳥、あるいは鳥から人へ感染する伝染病であり、人間から人間に感染する「インフルエンザ」とは区別されている。しかし「鳥インフルエンザ」は最近出現した語句であるため、単単語として辞書登録されていることが期待できない。この場合、形態素解析の単位を Mining Concept とする従来手法によると、この文書中には「鳥インフルエンザ」の事を話題にしている文書が多数存在するにもかかわらず、「インフルエンザ」についての話題が出現していると誤って分析してしまうことになる。これは分析の単位を間違って捉えているからであり、正しく分析するためには、「鳥インフルエンザ」を Mining Concept とすることが必要になる。

このように、分析の単位である Mining Concept を決定することはテキストマイニングの分析の精度を向上するために非常に重要である。我々は、話題の盛り上がりが時間的に変化する Mining Concept は分析上重要であると考え、Mining Concept の一種として話題における単語頻度の時系列変化を考慮するものを提案する。我々はこのような Mining Concept の一種を、時系列の頻度変化を考慮に入れていることから**時系列発生的** *Mining Concept (Fluctuation of Occurrence based MIning Concept: FOMIC)* と呼ぶ。

我々は FOMIC は複数の単単語から形成され、「A の B」のような 2 以上の名詞句、あるいは「A が B する」のような名詞とサ変動詞の組み合わせ（フレーズ）と言い換え可能である、と考えている。上記の例では「鳥」と「インフルエンザ」から成り立っており「『鳥』の『インフルエンザ』」であると説明できる。その概念の表現の一つが「鳥インフルエンザ」という複合語である。

また、FOMIC を構成する語は、時系列的に変化する同じ話題から出現していると考えられる。そのため、各々の名詞に関して、出現の時系列変化は同期していると考えられる。このアイデアに基き、我々は FOMIC らしさの指標を定義する。まず、文書中に出現する複合語を、複数の単単語から成り立つ概念の代表であると考え、FOMIC の候補とする。そして、その複合語を構成する名詞の中で重要な名詞と、他の単単語の出現の時系列変化が同期しているとき、その概念を FOMIC として扱うと定義した。

本論文の構成は以下の通りである。まず第 2 章において関連研究を述べ、第 3 章において本論文の目的であるテキスト分析の単位としての FOMIC について述べ、実際のデータを元に具体例を示す。第 4 章において我々の考える FOMIC の抽出手法について述べ、第 5 章、第 6 章においてその手法を用いた実験結果と考察を示す。

[1] ここでいう出現回数は、「鳥インフルエンザ」中の「インフルエンザ」も数えている。

2　関連研究

時系列に従った話題語抽出の研究は今までに複数提案されている。

時系列文書データからの重要語抽出については、Kleinberg らの話題を含む文書や固有名詞がその話題の起こった時間付近に集中して出現することを確率モデル中の異常値として表す研究 (Kleinberg, 02) が挙げられる。佐藤らは、まず文書を話題ごとにクラスタリングし、そのクラスタごとに時系列を用いて話題の広がりと伸びを定義し、特徴的な語を抽出することによって話題を表す語を抽出する手法 (Sato, 05) を提案している。

関口らは、ブログなどの筆者が特定できる文書から各筆者間の話題の類似度を求め、類似している人が共通に用いている語を話題語句として用いる手法 (Sekiguchi, 05) を提案している。

これらの研究はすべて、語の単位は認識されており、それの時間的推移や属性の違いによって重要度がつけられるものであった。語の単位を決定する複合語抽出の研究は、Ananiadou らによる C-Value/NC-Value を用いたもの (Ananiadou, 94) や、中川らによる複合語の出現頻度とそれに隣接する語の頻度に基づく専門用語の抽出 (Nakagawa, 03),(中川, 03) などがある。このように既存の複合語抽出の研究の多くはコーパス全体に関しての出現確率等を用いている。

このように、話題語抽出と複合語抽出には様々な研究が存在するが、抽出した語をテキスト分析の単位として利用するといった考察がされているものは少ない。我々は、複合語をテキストマイニングの単位である FOMIC の候補とし、更にその複合語を構成する名詞の時系列的な出現頻度の同期性を観測することにより、テキスト分析の単位である FOMIC らしさの指標をつけることを提案する。

3　FOMIC の抽出

この章では本稿における研究の目的において述べる。

3.1　FOMIC の定義

テキスト分析の主な目的として、新奇性の発見が挙げられる。新奇性とは、時間による概念の変動や、ある概念とある概念の出現の相関などである。ここで言う概念が、テキストを分析する際の単位である Mining Concept に相当する。

新しい概念を表すための新しい言葉は多くの場合、既存の言葉を組み合わせることによって作られる。たとえば、「エコ」と「家電」を組み合わせることによって、環境に配慮した家電を「エコ家電」と表現する。従って、昨今の環境関連のテキストを分析するためには、「エコ家電」を Mining Concept とする必要がある。

しかし、すべての Mining Concept が「エコ家電」のように複合語となって文書中に出現するわけではない。例として、「個人情報保護」という語が Mining Concept であるときを述べる。この Mining Concept は文書中に複合語として出現する場合のほかに、「個人情報の保護」や「個人の情報を保護する」のような名詞と名詞あるいは名詞とサ変動詞として現れる。しかしこれらは同じ概念を表しており、分析する際には一つの Mining Concept として扱うべきである。

これらの表現を一つにまとめるためには、人手で作成したパターン辞書を使うことが考えられる。そのためにはまず、文書内に出現している概念のうち、何が Mining Concept なのか、という定義が必要になる。

テキストマイニングの重要な分析として、大量の文書の中からの既存の知識と関連の深い話題の抽出や、その話題と関連する事象の抽出がある。我々は、話題とは時間的推移のあるものだと考えた。従って、単語もある話題を表現しているものであれば時間的推移を観測できると考えられる。

そこで、我々は単語の時間発生的推移を考慮した Mining Concept である FOMIC を定義した。FOMIC とは、対象とする文書内において以下の 2 つを満たすものである。

- Mining Concept を構成する名詞が隣接した複合語を構成し、出現していること
- Mining Concept を構成する名詞の出現頻度の時間的推移が同期していること

次の節ではこの定義に基いた FOMIC の抽出手法について述べる。

3.2　FOMIC 抽出の概要

我々が取り出したい FOMIC は、時間的に推移する話題と密接に関係するものである。この話題の時間的推移と、FOMIC を構成する名詞の出現頻度の時間的推移は同期していると考えられる。また、定義により FOMIC は複合語として分析対象の文書の中に出現している。

このアイデアに基づき実際のコーパスから FOMIC の候補を抽出する。まず、単単語で分析の対象とする文書中の話題の中心となっているものを抽出する。我々はこのために、長野ら (Nagano, 01) のアイデアを採用した。彼らのアイデアは、出現頻度の高い単語は一般的であるために、

また出現頻度の低い単語も特殊であるために文書内での情報量が少ない、というものである。従って、中頻度の単語は話題について重要な単語であると考えられる。これを**重要語**と呼ぶ。実際には、重要語として文全体から名詞（一般名詞・固有名詞）の単単語を取り出し、その頻度の中で中頻度のものを抽出した。

次のステップとして、その重要語を含む複合名詞を取り出す。Mining Concept は複数の単単語から構成されるが、我々は FOMIC は文書中で隣接した単単語、つまり複合名詞として出現していると仮定しているため、この複合名詞を FOMIC の候補とする。そして、その複合語を構成する単単語の頻度の時間推移を測定する。複合語中の重要語と、その他の単単語の頻度の時間推移が同期している複合語を取り出す。この複合語が表す概念が、FOMIC であるといえる。

3.3 FOMIC の具体例

我々の手法では、候補として抽出された複合語に対して FOMIC らしさの指標を与えることができる。その指標によって、FOMIC を構成する単単語の選択や数の考察を行うことができる。ここでは、我々の指標によってどのような FOMIC が適当とされるのか、具体例を示す。

3.3.1 FOMIC を構成する単単語数の判定

構成する単単語が多い Mining Concept は情報量が多いが、その一方でその Mining Concept を含む文書の頻度が少なくなると考えられる。頻度が少ないと分析を行う単位としては不適当である。従って、Mining Concept を構成する単単語の数を考察を行う必要がる。

我々の提案した FOMIC を抽出する手法を用いることにより、Mining Concept の単語数を決定することが可能である。「鳥インフルエンザ問題」という文字列が観測されたとする。このなかの「インフルエンザ」は重要語であるとする。すると、この文字列から考えられる Mining Concept の候補は

鳥　インフルエンザ　問題
鳥　インフルエンザ
インフルエンザ　問題

の 3 つである。ここで、先ほどの文書におけるこの複合語を構成する単単語のそれぞれの頻度の時間推移を見る。

図 2 を見て分るように、重要語である「インフルエンザ」と「鳥」は出現頻度の時間推移が似ているが、「インフルエンザ」と「問題」の出現頻度の時間推移にはほとんど相関はない。これは「鳥」と「インフルエンザ」は話題に時

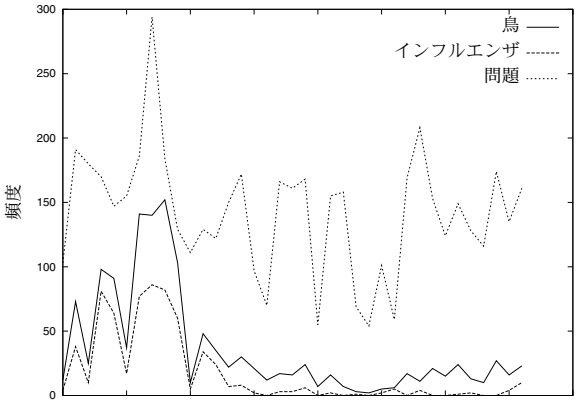

図 2: 「鳥」「インフルエンザ」「問題」の頻度の時間推移

期性がある（このコーパスは 1 月から 12 月までであるから、1 月から 3 月頃に話題の盛り上がりがある）のに対し、「問題」には時期性がないためである。そのため、Mining Concept としては「鳥インフルエンザ問題」や「インフルエンザ問題」とするより「鳥インフルエンザ」とするほうがより適切であるといえる。

3.3.2 FOMIC を構成する単単語の判定

FOMIC 候補は一つの重要語に対し、複数存在する。その複数ある候補について、どの構成単語を選択すると FOMIC らしいのか考察を行う必要がある。

例として、重要語である「自衛隊」を一部分に含む「自衛隊派遣」、「自衛隊撤退」、「自衛隊参加」の 3 つの複合語について考察する。これら 3 つの複合語の頻度の時間推移を図 3 に示す。

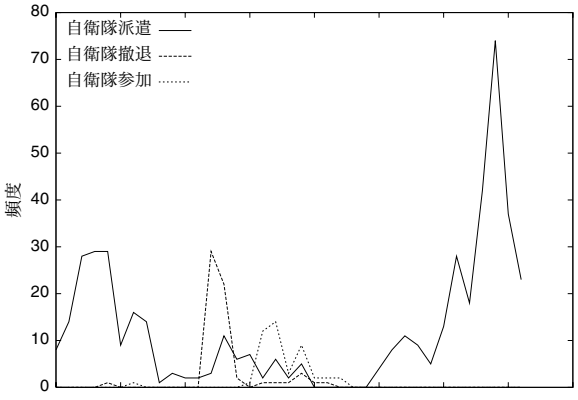

図 3: 「自衛隊派遣」、「自衛隊撤退」、「自衛隊参加」の頻度の時間推移

時系列推移を見ると、3つとも頻度の変化にピークがあるため、その日付の近辺に関連する話題であると仮定できる。通常の話題抽出の手法などを利用すると、3つの複合語がともに Mining Concept として抽出される可能性もある。そこで、この3つの複合語内に含まれる重要語（「自衛隊」）と、単単語（「派遣」「撤退」「参加」）重要語の出現頻度が同期しているか、1年にわたる時間推移を見てみる。

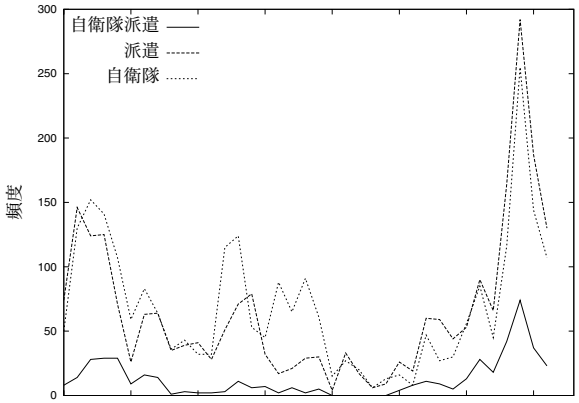

図 4: 「自衛隊」「派遣」の頻度の時間推移

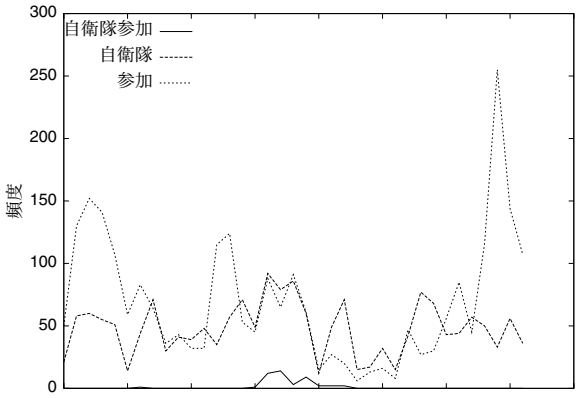

図 5: 「自衛隊」「参加」の頻度の時間推移

図 4 を見ると1年間を通し「自衛隊」と「派遣」の出現頻度の時間推移が同期しており、この2語は1年にわたって強く関係しているといえる。

図 5 と図 6 を見ると、それぞれ一時期のみ出現頻度の時間推移の同期が見られる。これらの期間ではそれぞれ「自衛隊」と「参加」、「自衛隊」と「撤退」は強く相関しており、それ以外では相関はない。

この結果から、分析対象が1年間であれば「自衛隊派遣」が FOMIC としてふさわしい。しかし、出現頻度の時間推移の同期性は分析を行う期間に拠るため、FOMIC は分析期間によって変化する。上の例において、短い期間の分析であれば「自衛隊撤退」や「自衛隊参加」も FOMIC らしさが大きくなることがあるといえる。

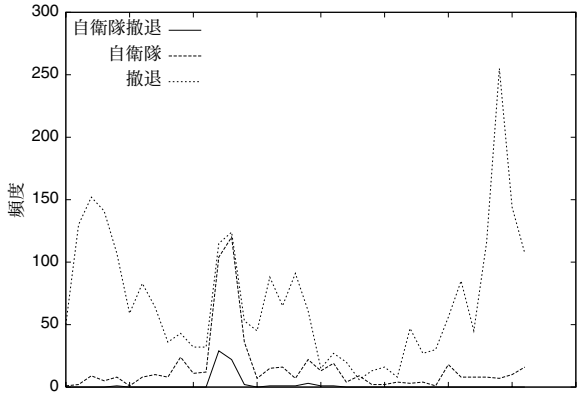

図 6: 「「自衛隊」「撤退」の頻度の時間推移

4 抽出手法

FOMIC 抽出には以下のステップがある。

- 重要語の抽出
- 得られた重要語を含む複合語の抽出
- 各複合語に対し FOMIC らしさの指標を付与する

ここでは各ステップの詳細について述べる

4.1 重要語の抽出

話題を表す名詞の重要語として、中頻度の名詞を抽出する。中頻度の名詞 $N\%$ を抽出するために、全体の名詞数の高頻度と低頻度から $\frac{100-N}{2}\%$ ずつ取り除く。この結果得られた名詞を重要語とする。

4.2 FOMIC の候補である複合語抽出

FOMIC の候補として、得られた重要語を含む複合語を分析対象の文書から取り出す。重要語が2つ以上含まれる複合語では、出現する重要語のなかで最後の単語を重要語とする。これは、日本語の場合後ろの単語が主辞になりやすいため、重要語の中で語順が後ろのものが話題と関連が深いと考えられるためである。

4.3 FOMIC らしさの指標の付与

m 個の単単語からなる FOMIC 候補

$$w_{all} = w_1 w_2 \cdots w_m$$

を評価する。この中での重要語を w_{core} とする。

まず、2 つの単単語の出現頻度の時間推移の差を定義するため、時刻の単位 ΔT を定める。時刻 t から ΔT の間に出現した単語 w の頻度を $f(w,t)$ と定義する。時刻 t_k と t_{k+1} における単語 w_i の出現頻度の差分を $\Delta f(w_i, t_k)$ とすると

$$\Delta f(w_i, t_k) = f(w_i, t_{k+1}) - f(w_i, t_k)$$

と表される。このとき、時刻 t_k における単語 w_i と単語 w_j の頻度の差分の差 $D_t(w_i, w_j, t_k)$ を定義する。

$$D_t(w_i, w_j, t_k) \stackrel{\text{def}}{=} \frac{1}{\Delta T} |\Delta f(w_i, t_k) - \Delta f(w_j, t_k)|$$

これを FOMIC を求めたい全区間（t_0 から t_{n-1} まで）足し合わせることで、単語 w_i と単語 w_j の頻度の時間推移の相違度 $D_T(w_i, w_j)$ が定義できる。

$$D_T(w_i, w_j) \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} D_t(w_i, w_j, t_k)$$

2 つの単単語の出現頻度の時間推移の相違度 $D_T(w_i, w_j)$ を用いて、FOMIC 候補 w_{all} に対し重要語と他の単単語との相違度をあらわす D_{all} を求める。このとき、単語数 $m-1$（重要語は除外する）で正規化を行う。

$$D_{all} = \frac{\sum_{i=1, i \neq core}^{m} D_T(w_i, w_{core})}{m-1}$$

この D_{all} が FOMIC らしさの指標となる。D_{all} が小さければ、複合語中の重要語と他の単単語の頻度の時間推移が同期していることになり、FOMIC らしいといえる。

相違度 D_{all} を計算をする際の単単語として、接尾辞・接頭辞は除外して計算した。これらは、頻度の時間推移が一般的すぎると考えられる一方で、話題をよく表すほかの名詞等と強く結びつき、場合によっては一つの単単語と考えたほうがいい場合もあるからである。

5 実験及び結果

本手法の有効性を調べるため、定義された尺度を元に、実際のデータで複合語を取りだし、FOMIC らしさの評価を行った。対象としたのはインターネット上での 2003 年一年分のデータのうち作成された日時が判別できたもので、その中から頻度 5 以上の複合名詞を対象とした。

重要語には、全名詞（単単語）のうち中頻度 10% を用いた。以下にその例を下に示す。

重要語の例 暫定、課題、義務、批判、先発、比例、時期、生産、システム、W杯、…

その重要語を含む複合語を、FOMIC 候補とした。

その複合語の FOMIC らしさの指標として、前章で定義した D_{all} を、$\Delta T = 7$ 日 として計算した。

FOMIC らしさの評価は、D_{all} の昇順に並べて閾値を設けてそれ以下の候補を検証する方法と、含む重要語ごとに分けられた候補の D_{all} の大きさを比べる方法とがある。

D_{all} は単語の頻度変化の大きさによって変化するので、すべての単語を同じ閾値で扱うのは適していないと考えられる。しかし、同じ重要語を含んでいるものは同じ話題と結束性が高いと考えられるため、比較するのに適していると考えられる。そのため、本手法では重要語ごとに分けられた FOMIC 候補を検証する方法を採用する。

以下は「財産権[2]」「高齢者[3]」「インフルエンザ[4]」を重要語とする例である。

Mining Concept 候補	D_{all}
知的 財産権	9.00
知的 財産権 保護	59.50
財産権 保護	107.00

表 1: 「財産権」を含む FOMIC 候補

Mining Concept 候補	D_{all}
高齢者 介護	19.00
高齢者 世帯	21.00
高齢者 社会	230.00

表 2: 「高齢者」を含む FOMIC 候補

Mining Concept 候補	D_{all}
鳥 インフルエンザ	36.00
鳥 インフルエンザ ウイルス	88.50
鳥 インフルエンザ 発生	99.50
鳥 インフルエンザ 対策	131.50
鳥 インフルエンザ 対策本部	142.00
鳥 インフルエンザ 事件	324.50
鳥 インフルエンザ 問題	367.00

表 3: 「インフルエンザ」を含む FOMIC 候補

[2] 「権」は接尾辞であるため一つの単単語としては扱わず、計算には含めなかった。

[3] 「者」も接尾辞であるため、計算には含めなかった。

[4] 文書中には「鳥インフルエンザ」ではない「インフルエンザ」の話題がほぼ見られないため、ここでは「鳥」を含んでいないものは載せていない。

6 考察

テキスト分析の目的は新奇性の発見であることはすでに述べた。従って、その分析の単位である Mining Concept には、検索や用語抽出と違い正解を定めることができず、抽出手法に対し精度を測るということが行えない。

そこで、我々は抽出した FOMIC に対する評価として、抽出した FOMIC がテキストマイニングの結果に与える影響を考察した。そのために、Mining Concept を含む文書に絞り込み、頻度順に並べた名詞のリストを見比べてみる。もし、その Mining Concept が文書集合において特徴のある文書を絞り込むことができるのであれば、名詞の頻度順も変化するはずである。このとき、その Mining Concept は、分析をする際に新奇性のあるものを発見できる一つの性質を持っているといえる。この方法で、選ばれた FOMIC がどの程度 Mining Concept らしいか、考察を行う。

表 3 より、我々の手法で与えた FOMIC、つまり Mining Concept らしさは「鳥インフルエンザウイルス」「鳥インフルエンザ発生」「鳥インフルエンザ問題」の順であった。表 4 にそれぞれの Mining Concept で絞り込んだ時の名詞の頻度の高いものから上位 30 位までを示す[5] 。下線を引いてある名詞は、「鳥インフルエンザ」で絞り込んだ時の順位が 30 位以内に現れていなかった名詞である。また、表 5 にそれぞれの Mining Concept における上位 30 位内における、「鳥インフルエンザ」で絞り込んだ時の順位が 30 位以内に現れていなかった名詞の数と割合を示す。

	ウイルス	発生	問題
30 位以内に入った単語の個数	22	19	14
30 位以内に入った単語の割合	73%	63%	47%

表 5: Mining Concept による名詞分布の変化

表 5 からわかるように、「鳥インフルエンザウイルス」や「鳥インフルエンザ発生」を用いて文書を絞り込んだ時は、「鳥インフルエンザ問題」で絞り込むよりも名詞の順位の変化が大きい。つまり、Mining Concept らしいということが言え、我々の FOMIC らしさの指標と合っている。この評価に基くと、我々の求めた FOMIC の指標は Mining Concept らしさの指標と一致するといえる。

7 まとめ

我々は、テキストマイニングの分析単位である Mining Concept の一種である FOMIC を定義し、抽出する手法を提案した。この手法は、複合語中の各単単語が同じ話題に属しているかどうかを、単語の頻度の時間推移の同期性によって捉えるものである。我々は FOMIC らしさの指標を定義し、FOMIC 候補を評価した。また、得られた FOMIC らしさの指標が Mining Concept らしさと一致しているかを、テキストマイニングの一手法である文書集合中の名詞分布の観測という方法で考察し、我々の提案した手法が妥当であると確認した。

実際に Mining Concept を決定するには、我々が提案したような指標を用いて人による判断が必要になる。そのため、文書に対し重要な概念から判断する必要がある。今回は頻度の時間推移のみを考慮に入れたが、実際の分析への影響を考えると、全体に対する出現頻度の割合や複合語の長さも考慮に入れなくてはいけないだろう。

今回、我々が取り出した FOMIC は、「ＡのＢ」や「ＡがＢする」といった名詞句やフレーズに言い換え可能な概念であった。我々の手法では単単語での出現頻度の時間変化を見ているため、複合語中のみならずこのような表現中の単単語も捉えることができる。しかし、実際に「ＡＢ」という FOMIC が含まれているかどうかを文書から得るためには、フレーズの形である「ＡのＢ」や「ＡがＢする」といった表現まで捉える必要がある。そのため、Mining Concept を文書中で認識するためには、フレーズ単位での同義性の吸収が必要である。

我々はこの論文において Mining Concept のテキストマイニングでの有効性を述べたが、通常の複合語抽出と違い評価が困難であるのが現状である。テキストマイニングの結果は、Mining Concept の選択に影響を受けるため、テキストマイニングの評価は Mining Concept の評価ともいえる。Mining Concept の抽出における評価について指標を決める必要があり、今後の課題であるといえる。

[5] 名詞には複合名詞も含む。

References

S. Ananiadou 1994. “A Methodology For Automatic Term Recognition”. *COLING 1994: pp. 1034-1038*

J. Kleinberg 2002. “Bursty and Hierarchical Structure in Streams,” *In Proc. of KDD 2002, pp. 91-101*

佐藤吉秀, 川島晴美, 佐々木勉, 奥雅博. 2005. 時系列ニュースにおける最新話題語抽出方法. **情報処理学会自然言語処理研究会** *NL168, pp. 1-12*

関口裕一郎, 佐藤吉秀, 川島晴美, 奥田英範, 奥雅博. 2005. blog ページ集合に対する話題語句抽出手法. **情報処理学会自然言語処理研究会** *NL170, pp. 27-32*

Nagano T., Takeda K. and Nasukawa T. 2001. “Knowledge Discovery using Robust Natural Language Processing” *In Proc. of PACLING 2001*

	鳥インフルエンザ	鳥インフルエンザウイルス	鳥インフルエンザ発生	鳥インフルエンザ問題
1	鳥取県	鳥小屋	鳥類	鳥小屋
2	インフルエンザウイルス	インフルエンザウイルス	インフルエンザウイルス	野鳥
3	感染経路	ウイルス感染	罰金	インフルエンザウイルス
4	こども	きぐ	山口市	農水省
5	鶏舎	感染者	農水省	京都府警
6	農水省	羽毛	コネティカット州	鶏舎
7	潜伏期間	禁止措置	野鳥	担当者
8	夜	農水省	養鶏農家	船井
9	発売	炭谷	農相	農場
10	京都府警	野鳥	輸入再開	感染経路
11	養鶏業者	鶏肉	鶏肉	措置
12	羽毛	マニラ	鶏肉加工	養鶏農家
13	農場	死者	中国産	農政部
14	前	船井	環境省	山口市
15	中国産	全域	１２月	反応
16	山口市	農場	京都府警	府警
17	鶏肉加工	養鶏農家	タクシン	羽毛
18	農政部	ハノイ	パキスタン	刑事告発
19	タイプ	検出	葵	ウイルス感染
20	パキスタン	人間	違反	違反
21	ウイルス感染	農政部	日本国内	保健衛生
22	輸入禁止	Ｈ５Ｎ１	２月	１２月
23	日本リサーチセンター	韓国政府	訓子	カラチ
24	ベトナム政府	敷地	船井	作業服
25	カラチ	血液	地元	大量
26	義務	山林	立ち入り検査	家庭
27	家庭	タイプ	山口県	鶏肉加工
28	府警	家	大量死	陸上自衛隊
29	決議	住民	農場	チーム
30	人間	反応	羽毛	デオキシリボ核酸

表 4: Mining Concept による名詞分布への影響

Nakagawa H. and Mori T. 2003. “Automatic Term Recognition based on Statistics of Compound Nouns and their Components” *Terminology, Vol.9 No.2, pp. 201-219*

中川裕志、森辰則、湯本紘彰. 2003. 出現頻度と連接頻度に基づく専門用語抽出. **自然言語処理**、*Vol.10 No.1, pp. 27-45*

Nasukawa T. and Nagano, T. 2001. “Text analysis and knowledge mining system.” In *IBM Systems Journal, Vol.40, No.4, pp. 967-984*