原著論文

未来の流行を予測する Web からの 注目キーワードの発見†

砂山 渡* 谷内田 正彦*

WWW 上には多くのリソースがあるため、それらから有用な情報を抽出することはごく当たり前のことになってきた。本稿では、注目キーワードを抽出するシステムを提案する。注目キーワードはユーザの興味に関連し Web ページ中に新たに出現した単語である。すなわち、観点と時間が重要なファクターとなる。ある話題の主題は観点によって異なり、世界は時間によって変化する。このように、注目キーワードは Web ページに表れる大衆のリアルタイムな意見を代表し、ユーザの観点を反映するため、未来の流行予測に役立つ。そのため、創造的なユーザは出力される注目キーワードを自身の活動に適用できる。

キーワード：流行、注目キーワード、WWW、サーチエンジン

1. はじめに

WWW は情報の宝庫である。ニュース、天気予報から最新の研究の動向に至るまで、ありとあらゆる情報が備わっている。これらの情報の中から世の中のニーズや流行をいち早く捉え、新しい研究や商品開発を始めることは、社会における成功をおさめるためには必要なことである。しかし、インターネットユーザにとって必要もしくは重要な情報を取り出す操作は情報量の増加にともなって困難になってきた。現在、インターネット上の情報を探す手段としては、サーチエンジン（検索エンジン）が広く用いられており、検索の効率を上げるために、サーチエンジン自体の性能向上や、関連フィードバック [Rocchio 71] や検索式拡張 [Beaulieu 97] などによる再検索支援といった研究が数多く行なわれている。数多く行なわれているということは、それだけの需要および効果があるということに加え、そこに一つの流行が存在すると捉えられ、同様に Web 上のデータにも世の中のニーズや個人の興味を反映したいくつかの流行が潜んでいる。そこで、本研究においては世の中のデータの移り変わりを反映した、将来の流行を予測できるキーワードを、Web 上のデータから取り出して提供することを目指す。

キーワードを提供するということは通常の検索とは異なり、文章（Web ページ）ではなく単語（キーワード）で結果を返す。たとえユーザの興味に関連する多

くの文章を検索で得られたとしても、それらのページ全体を見ながら共通する観点を発見したり、最近の新しい傾向を発見することは困難であると同時に多大な労力を要する。しかし、キーワードの使われている具体的な文脈を参照することは有益であるため、この文章と単語の2種類の出力は用途によって使い分けられることが必要となる。本研究では、世の中の流行のきっかけとなり得るキーワードと、そのキーワードが実際に使われている文脈とを併せて提供するシステムの構築を目指す。

存在する多くの情報の間の関係を決定する要素として「観点」と「時間」とが挙げられる。観点とは、見方によって情報の主題が変わるということであり、たとえば、アイドルが始球式を行なったというニュースは「アイドル」という観点に基づけば芸能ニュースとして芸能との関連が強くなり「野球」という観点に基づけばスポーツニュースとしてスポーツに分類される [笠原 94]。また世の中は常に変化、変遷しているという事実から、情報の発信時期の情報は、各時点における世の中を知るために有効に役立てられる [小池 97]。たとえば、過去に二つの事象 A と B の間に強い相関があったとしても、現在そのような事実はない場合、過去の知識と現在の知識は区別されるべきである。そこで本研究においては、これら観点と時間とを組み合わせ、ユーザの入力する検索語を観点としたキーワードを見つけるとともに、Web ページの最終更新日時を情報の発生日時として、情報をその発信時期によって分類した上で、ユーザが指定する時期に特有のキーワードを抽出する。本稿では、これら2つの要素を組み合わせた注目キーワードおよび、注目キーワードを含む重要文を出力し、人間が未来の流行を予測できる環

† Discovery of Notable Keywords from Web Pages for Estimating Future Trends

Wataru SUNAYAMA, Masahiko YACHIDA

* 大阪大学大学院基礎工学研究科システム創成専攻
Dept. of Systems Innovation, Graduate School of Engineering Science, Osaka University

境の構築を目指す。

以下、2章で研究の位置付けを明確にし、3章で注目キーワード抽出のアルゴリズムを説明し、4章で構築したシステムの実験を行ない、5章で結論を述べる。

2. 研究の位置付け

まず、注目キーワードの抽出に関して、膨大なデータの中から有効な知識を取り出す研究は、データマイニング [Fayyad 96、福田 01] の分野で盛んに行なわれている。本研究もこの流れに沿ったものであり、いかにして「有効さ」を定義して情報を絞り込めるかが最大の課題である。しかし、真に知的な作業を行なうのは人間であり、有効な情報を見つけるという作業の知的レベルは高い。データマイニングで抽出されるルールも、最終的な選別は人間が行なっているのが現状である。本研究は、データマイニングを行なう前処理の段階として位置付けられる。データマイニングでは、知識発見に有効であると考えられる範囲にまでデータを絞り込んでからマイニング処理を行なっているが、本研究の出発点はユーザの興味に関連する広範囲のデータである。世の中から実際のチャンスを発見するためには、広範囲のデータをデータマイニングの処理が行なえるまでに絞り込む操作が大きなウエイトを占める。これは Genetic Algorithm (GA) [Goldberg 89] において各染色体の質が最終的に得られる解の質に関係するのと同様に、同じデータマイニングの手法を用いる場合においても、もとデータの質が高い方が良いと考えられる。したがって本研究は、広範囲のさまざまな Web 上のデータの中から、世の中の流行に関わる質の良いデータの提供を目指している。

注目キーワードの2つの要素の一つである「観点」に関して、Web からのキーワード抽出を行なう研究の多くは、再検索支援やテキスト分類を目的としているため、単純に多くのテキストに共通の単語をキーワードとしている。そのため、世の中のチャンスを発見するという目的に際して、高頻度ではない低頻度の単語をキーワードとして積極的に捉える新しい枠組が必要となる。そのため、次章で述べる展望台システム [砂山 02] の出力する特徴キーワードをベースに用いる。特徴キーワードは、ユーザの観点にのみ関わって出現する単語であると同時に、その抽出アルゴリズムは低頻度の単語ほど優先して出力する。

注目キーワードの要素である「時間」に関して、過去のデータから未来を正確に予測することは極めて難しい。最も卑近な未来予測に天気予報があるが、これも過去の統計データに基づく確率的な予測であり外れることも多い。Web 上のデータは時々刻々と変化して

おり、[小池 97]のように過去の Web データ全体を意図的に保持しない限り、過去のデータにお目にかかることすら難しい。また、そのようなデータを保持することも、Web 全体の容量から考えると容易とはいえない。過去のデータがない場合に問題の1つに、削除されたデータが何であるかわからないという問題があるが、今後の流行を探るという目的に対しては、廃れていくデータは直接的には意味をなさないと考えられる。また、本研究では現存する Web ページのみから、各 Web ページの最終更新日時を扱うため、情報出現の日はあくまでも近似となる。しかし、1つの Web ページに載っている情報量はそれほど多くないことや、更新が行なわれた時点においては、その内容全体は有効かつ意味をなしていたと考えられる。いずれにしても、未来予測としてシステムができることは「あたり」をつけることであり、本研究では、最終的に有効なものはユーザに選びとってもらうことを前提に、流行の必要条件として世の中のニーズが増してきていると判断されるキーワードを抽出する。

3. 注目キーワードの抽出アルゴリズム

本章では、多くの Web ページの中から流行のきっかけとなり得る、注目キーワードを抽出するアルゴリズムについて述べる。

3.1 キーワード抽出アルゴリズム

0. データベース構築：Web ページをその最終更新日時と共にダウンロードする。各ページからキーワード候補を抽出する。キーワードの候補としては、形態素解析 [Chasen] によって、名詞^{*1}を取り出す。
1. ユーザからの検索語を入力（検索条件）として受け取る。
2. サーチエンジンによって、検索条件に合う Web ページ集合 (G) を獲得する。
3. 注目キーワードを得たい時期 (T) の入力をユーザから受け取る。ただし、入力がない場合は最新の6ヵ月とする。
4. G を入力を受けた時期 T に更新された Web ページ集合 G_{in} と、それ以外の時期の集合 G_{out} とに分割する。
5. G_{in} と G_{out} のそれぞれから、サーチエンジンによるランキング上位の Web ページを一定数 (デフォルトでは20ページ) 取り出し、3.3節で後述す

*1 形態素解析の弱点である平仮名の切り分け失敗に伴って現れる、一文字の平仮名名詞はノイズとして除去している。

る展望台システムによって、各 Web ページに含まれるキーワード候補 w に、観点キーワードとしての評価値と特徴キーワードとしての評価値を与える。

6. 単語 w の G_{in} における特徴キーワードとしての評価値から、 G_{out} における特徴キーワードとしての評価値を引いたスコアを、単語 w の注目キーワードとしての評価値 $Note(w)$ とする。詳細は 3.3 節で後述する。
7. 評価値 $Note(w)$ の高いキーワード w を検索語として、サーチエンジンを用いて各月に更新された Web ページ数を計算し、時期 T に更新された Web ページ数がそれ以外の時期に更新された Web ページ数より多いキーワード w の評価値 $Note(w)$ を上げ、逆に Web ページ数が減っているキーワードの評価値を下げる。詳細は 3.4 節で後述する。
8. 評価値 $Note(w)$ の高い順に注目キーワードとして出力する。

3.2 キーワードからの流行予測

本節では、システムが出力するキーワードと流行予測との関係について具体的に述べる。そのためにまず、流行ということばを「世の中に新たに広く認知されること」として定義する。すなわち、本システムが出力する注目キーワードは、現在の世の中ではまだあまり知られていないけれども、これから広く使われることになるだろう単語として意味づけされる。

「新たに」という説明は必ずしも必要ではないが、世の中に先駆けるという意味において、流行の後追いではなく流行を創るためには欠かせない要素である。本システムでは注目する時期を入力として与える必要があり、流行予測のために「最近数ヵ月」を入力として与えることを想定している。

流行予測は研究者が新規テーマを設定することに喩えられる。新規性のない研究は世の中で受け入れられないため、研究者は誰よりも早く新しい可能性に気づく必要がある。新しい研究の始め方には次のものが考えられる。

1. 研究者自ら新規なテーマを発掘する。Web 上でまだ誰も実現していない斬新な新テーマを自ら提案する。
 2. 少し流行の兆しが見え始めたテーマの流れに乗って、そのテーマを越えるあるいは派生する新たなテーマを設定する。
1. については、新テーマが世の中に受け入れられる保証がないため、成功した際には先駆者としての榮譽

が得られる反面、失敗した時には少なからず無駄な時間を過ごすリスクがある。2. は世の中の流れに沿っているため、ニーズも高く、研究としてのやりがいがある反面、他の研究者との競争に勝ち、競争を避ける必要がある。研究者がこの 1. と 2. のどちらを重視するかによって、最近 1 ヶ月に急に現れた話題に着目するのか、ここ 1 年流行となっている話題が欲しいのかが異なってくる。

すなわち、この入力される期間の長さがチャンスとリスクのトレードオフであり、期間が短いほど新規性が高く大きなチャンスを掴む可能性がある反面、失敗に終わるリスクも高確率で背負うことになる。逆に期間を長くするにつれて情報の信頼性が高くなるため、失敗のリスクが減る反面、チャンスとしての意味合いも薄れてしまう。現時点では、世の中に現れ始めて一般に情報が伝搬するのに十分な半年という期間をデフォルト値として設定しているが、この期間の設定にはさらなる検討が必要である。

次に「広く認知される」ことに関して、本システムが出力する注目キーワードは、少なくとも一つの Web ページにおいて検索語との関連が示されたキーワードである。ここで、すでに認知されてしまっている事柄を取り出しても、流行予測には意味をなさない。しかし、実際の効果として研究者を例に具体的説明を次に挙げる。

- a. Web ページの作者の多くは研究者ではない。そのため、はやりかけている内容をいち早く研究に採り入れるという点での効果が期待される。
- b. 単語の多義性により、入力した検索語と同じではない意味で使われた Web ページからも、キーワードが出力される可能性がある。同じ単語が使われる文脈には何らかの関連が存在すると考えられ、普段は捨てられる知識の中からの気づきを促し、新たなアナロジーを発見する可能性がある。

すなわち、流行予測には気づきが重要であり、そのきっかけとしての注目キーワードを本システムが出力する。

3.3 展望台システム

文章からキーワードを抽出するために用いる展望台システム [砂山 02] のアルゴリズムを説明する。展望台システムは重要文抽出システムであるが、重要文を抽出する際に文中で用いられる各単語を評価している。本研究においては展望台システムが出力する特徴キーワードをもとに、注目キーワードを抽出する。以下で観点キーワードと、特徴キーワードの定義および、そ

れらキーワードとしての評価値を計算するアルゴリズムについて説明する。

1. 観点キーワード：ユーザが頭に描いている自身の興味を表す単語、またその興味を表す単語と文章中で密接に関連する単語
2. 特徴キーワード：観点キーワードが出現する時のみ使われ、ユーザの興味をある方向へと特徴付ける単語

[展望台システムアルゴリズム]

ステップ1：観点キーワードの入力

ユーザに入力された検索語を観点キーワードの初期集合 S とする。

ステップ2：観点キーワードの補完

文章中の全ての単語に次の式による評価値 $view(w)$ を与え、その評価値による上位（単語種類数の4%で最大5個*2）に含まれかつ、まだ S に属さない単語を新たな観点キーワードとして S に加える。ただし、 $n(w)$ を単語 w が出現する文の数としている。

$$view(w) = \prod_{s \in S} \frac{n(w \wedge s)^2}{n(w)n(s)} \quad (1)$$

この式は単語 s が現れた時に単語 w が現れる条件付確率と、単語 w が現れた時に単語 s が現れる確率との積を表しており、二つの単語の共起頻度をその出現回数で正規化した値に相当している。このステップ2を新たな観点キーワードが追加されなくなるまで繰り返す。

ステップ3：特徴キーワードの発見

次に文章の観点となっている観点キーワードと強く関連して、文章を特徴付けているキーワードを抽出する。そのための単語 w の評価値 $feat(w)$ を式(2)で与える*3

$$feat(w) = \prod_{s \in S} \frac{n(w \wedge s)}{n(w)} \quad (2)$$

すなわち、観点キーワードが出現する時（文）にのみ出現する単語を探して、どれだけ多くの観点キーワードと密接に関係しているかを表す評価値を各単語に与える。この操作によって、特徴キーワードが現れる文には必ずいずれかの観点キーワードが存在する。

特徴キーワードは通常は無視されやすい頻度の低い単語を評価するために抽出されるキーワードであり、たとえ1度しか出現しない単語であっても、その単語が出現する文に観点キーワードが出現していれば評価される。そのため、単純に頻度が高い単語によるキーワード[Luhn 58]以外の独創的な単語を抽出するのに適している。また、入力した単語との関わりをもつという点において、他の文に現れない単語を取り出すTFIDF法[Salton 97]とも異なっている。すなわち、展望台システムが出力する特徴キーワードは、ユーザが意図的に入力した観点に関連して、現時点では公に現れていない低頻度の単語となるため、将来の流行予測に適したキーワードである。

3.4 時間情報とヒット件数

本研究で扱うのはWebページから取り出される「単語」であり、数多くの単語の中から有効と評価した単語をキーワードとして取り出す。これら単語の有効性を判定する材料の1つとして時間情報を用いる。すなわち、ある単語の特定の時期における出現の有無や頻度を調べることによって、その単語がどれほど世の中で用いられてきたか、これからも用いられそうかを判定する。また複数の単語に対しても、同時期にどれだけの単語が出現していたかを調べることによって、単語間の相関を知ることができる。

そこで、まず検索結果であるWebページ集合を、最終更新日時がユーザが注目する時期 T に含まれるWebページ集合 G_{in} と、それ以外の時期に含まれるWebページ集合 G_{out} とに分類する。この分類を元に、前節で求めた各Webページの特徴キーワードの評価値を用いて、各単語 w の注目キーワードとしての評価値を次式にしたがって計算する。

$$Relation(w) = \sum_{G_{in}} feat(w) - \sum_{G_{out}} feat(w) \quad (3)$$

すなわち、注目する時期 T にのみ出現する単語の評価値を上げ、それ以外の時期に出現する場合には評価値を下げる操作を行なっている。単語の評価値は展望台システムによって文章中での単語の役割を評価した値であるため、各Webページ中で意味や役割のない単語にはもともと評価値が存在せず、単純に単語の出現の有無によって差分をとった評価値とは異なる。

しかしこの評価値は、限定されたWebページ集合の中での評価値に過ぎないため、データベース全体での各単語の時間情報を考慮するために、各単語を検索語として検索した時のヒット件数（検索条件に適合するWebページの数）を評価に入れる。ヒット件数は世の

*2 5個という値は、文章の主題として論文に筆者が記すキーワードの数や、Webページ検索において特定のWebページを探すために用いられる検索語の数が、いずれも多くて5個程度であることより定めた。

*3 展望台システムは用途に応じて、特徴キーワードとして式(2)の値の大きい単語を一定数取り出すが、本稿で提案するシステムはこの式の値のみを利用する。

中の流行を反映することが多く、世の中で話題になるとそれに関連する Web ページが急増する。したがって、ヒット件数はそのまま流行のバロメータとして扱うことができる。しかし、Web ページの総数は日々増加するため、単純にヒット件数の増加を人気の増加と解釈することはできない。例えば、データベースのデータ総数の増加率を越えて増加する検索語は話題が高まっていると解釈できる。また同じ時期であれば、他の検索語によるヒット件数よりも多ければ、その検索語に比べると話題性が高いという相対的評価が可能である。

ヒット件数による単語の評価値 $New(w)$ を次のように定義する。ただし過去一年間の単語 w のヒット件数を $Hit(w, past)$, $past = -1, -2, \dots, -12$, ($Hit(w, -3)$ は 3 ヶ月前のヒット件数) とする。

$$Ave(w) = \sum_{k=-12}^{-1} Hit(w, k) / 12 \quad (4)$$

$$Wei(t) = \begin{cases} t + 6 & (-12 \leq t \leq -7) \\ t + 7 & (-6 \leq t \leq -1) \end{cases} \quad (5)$$

$$New(w) = \frac{\sum_{k=-12}^{-1} Hit(w, k) * Wei(k)}{Ave(w)} \quad (6)$$

すなわち、重み関数 $Wei(t)$ によって半年以上前に出現していれば負の評価値を、最近半年に出現していれば正の評価値を与えている。評価したい時期を変更したい場合には、この重み関数を好みに応じて変更することで可能である。またヒット数そのものよりも、ヒット数の増減の割合を重視するために、過去一年間のヒット数の平均で割って正規化を行なっている。

式(3)、式(6)を元に、最終的に出力される注目キーワードの評価値を式(7)で与える。

$$Note(w) = Relation(w) * New(w) \quad (7)$$

すなわち、ユーザの観点との関連度と情報の新規性との積として注目キーワードを定義し、ユーザの思考と世の中のデータとをつなぐことでユーザの興味に関わりかつ世の中の動向を捉えた流行の予測を支援する。

4. 実験システム

作成した注目キーワード抽出システムに関する実験を行なった。データベースは2002年1月にサーチエンジン yahoo [Yahoo] のページ検索を用いて、「テキスト」で検索した結果および「テキストマイニング」で検索した結果の合計1187ページである*4本実験では一度検索語で絞り込まれた Web ページ集合をデータベースとして扱っているが、実際のより大規模なデータベ

ースに対しても、一度検索語で絞り込んだ後の Web ページを対象とすることで、本実験と同様の効果が得られると考えている。

4.1 注目キーワード抽出システムの動作例

ある研究者が、最近のテキスト系の Web サイトにおける世の中の動向を調べ、創造的な研究を行ないたいと考えた場合を想定する。データベースに対して、検索語「創造」で検索を行なったところ60件の Web ページがヒットした。この結果を元に、出力された注目キーワードを表1に示す。表中の Note、New、Relation はそれぞれ注目キーワードとしての評価値(式(7))、時間情報による新規性(式(6))、ユーザの興味との関連度(式(3))の各値を表し、Hit が検索語とキーワード(Keyword)を共に含む Web ページの数を表している。

ユーザは自身の興味である創造に関連するキーワードを見て、自らの研究活動を世の中の動向と合わせて発展させるために適切なキーワードを選び、研究の方向性を検討する。ここで、「環境」「情報」「活動」など検索語「創造」との関連が理解できる語については、それらの単語が現在盛んになってきているという現状を踏まえ、その流れに乗って研究を行なっても良いし、何らかの形でこれらを研究に採り入れる工夫を行なうことができる。また、「フォーム」「快適」という出現頻度が低い単語に対して、「創造」という単語との組合せをすぐ正確に把握できない場合もある。これらの単語に関してはその出現理由を調べる材料として、検索語と注目キーワードを AND 条件として検索した結果(図1)を用いる。この図中の要約は検索結果の Web ページを展望台システム [砂山 02] によって要約した結果であり、観点キーワード(検索語)と各 Web ページ内における特徴キーワードを含む重要文となっている。ユーザはこれを参照することで、「モチーフとフォームの歩み/創造のシステム」「快適空間を創造」という表現を見出し、「創造」との関連を把握することができる。この、検索語と注目キーワードとの同時出現頻度が情報の信頼度に相当するため、信頼度が高い単語を使うことは安定した流行を追うことに、また信頼度が低い単語を使うことは新たな流行を作ることに当たる。

注目キーワードを新規性順に並べたものを表2に、関連度順に並べたものを表3に示す。また、図2に注目キーワードの評価値上位100語の、新規性と関連度の相関図を示す。図2の○印のついた単語は、評価値の上位として表1にある単語である。

*4 yahoo の検索結果は合計でおよそ100万件になるが、そのうちで yahoo の検索結果ページを元に実際に獲得可能であった Web ページは上記で全てであった。

表1 注目キーワード (Note 順)

Keyword	Note	New	Relation	Hit
フォルム	72	72	1.0	1
快適	67	45	1.5	2
環境	66	26	2.6	32
企業	64	22	2.9	32
ケミカル	60	60	1.0	1
テクノロジー	52	26	2.0	9
セルフ	49	49	1.0	2
テキスト	45	23	2.0	52
情報	44	24	1.9	45
対応	44	22	2.0	21
音楽	44	44	1.0	1
進歩	43	43	1.0	2
活動	42	23	1.8	23
スケール	41	41	1.0	1
解明	41	21	1.0	2

図2からわかるように、新規性と関連度の間の相関はなく、最終的に評価値上位として出力される単語には、新規性と関連度が共に高い単語が出力される。関連度が1.0の値が多いのは、検索語と同じ文中に現れる頻度が低い単語が数多く存在するためであり、このような頻度が低い単語からも時間情報を用いて出力が得られる^{*5}。しかし、時間情報のみを用いた表2においては新規性の高い単語が並んでいるが、頻度が低い単語ばかりとなって検索語との関連がわかりにくい単語が並び、いわゆるノイズとなる単語が多くなってしまふ。逆に、関連度のみが高い表3の単語では、「社会における創造」「創造による開発」といった、良い意味では定着した、悪い意味では使い古された頻度の高い単語が出力に多く含まれるため、世の中の新しい流れを捉えるには不向きであると考えられる。

4.2 注目キーワードの評価実験

システムが出力する注目キーワードを評価するための実験を行なった。しかし、ユーザが実際に流行予測を行なえるかという点に関しては、人間の知的能力に関わる部分であり、本システムはその知的作業を支援するという立場をとるために、流行予測のための材料として出力する注目キーワードがユーザの目に留まるか、ユーザに積極的に用いられるかという観点で実験を行なった。被験者は20代の研究者、大学院生および大学生の男女計20名で、実験手順は以下の通りである。

1. 2001年11月までのデータを用いて「創造」とい

*5 観点キーワードと同じ文中に現れる頻度が1回である単語には1.0以上の関連度の値が与えられないため、関連度1.0以上の単語を評価している。

図1 関連トピック検索結果

表2 注目キーワード (New 順)

Keyword	Note	New	Relation	Hit
フォルム	72	72	1.0	1
ケミカル	60	60	1.0	1
セルフ	49	49	1.0	2
快適	68	45	1.5	2
音楽	44	44	1.0	1
進歩	43	43	1.0	2
スケール	41	41	1.0	1
解明	41	41	1.0	2
リズム	39	39	1.0	2
ネーム	38	38	1.0	2
総合	38	38	1.0	18
家庭	37	37	1.0	4
独創	36	36	1.0	2
素材	34	34	1.0	7
浸透	34	34	1.0	2

う単語で検索を行ない、式(3)の評価値の上位100個のキーワードを獲得する。

2. 実験手順1.で獲得した100個の単語に次の5種類の手法による評価値の上位15個のキーワード集合をそれぞれ用意し、被験者に「独創的な創造を行なうために必要と思われるキーワードをそれぞれ3つ選んで下さい。」という質問をしてその解答を得る。この質問は、ユーザの自由な発想によって回答を分散させた中で、最も選ばれ

表 3 注目キーワード (Relation 順)

Keyword	Note	New	Relation	Hit
価値	29	9	3.2	12
企業	64	22	2.9	32
環境	66	26	2.6	32
システム	41	20	2.0	42
生活	38	19	2.0	8
テクノロジー	52	26	2.0	9
変化	38	19	2.0	9
発展	40	20	2.0	15
対応	44	22	2.0	21
テキスト	45	23	2.0	52
情報	44	24	1.9	45
活動	42	23	1.8	23
知恵	30	17	1.8	4
社会	35	21	1.7	27
快適	67	45	1.5	2

図 2 新規性と関連度の相関

やすいキーワードを確認するためのものである。比較対象は本稿で用いている 3 種類の評価値によるキーワードと、単純な頻度情報によるキーワード、また頻度情報による短期的な時間評価によるキーワードの 5 種類である。なお、各キーワードは表 1 から表 4 に示すキーワードとそれぞれ 6 ~ 9 割程度同じである。

○ Note：式 (7) の単語の注目度の評価値

○ New：式 (6) の単語の新規性の評価値

○ Relat.：式 (3) の単語の関連度の評価値

○ Freq.：単語の 2001 年 11 月の頻度*6

○ Diff.：単語の 2001 年 11 月と 12 月の頻度の差

これら 5 種類のキーワードの提示順序は無作為とする。

3. 被験者が実験手順 2. で選んだ 5 種類 × 3 個の合計 15 個のキーワード（重複がある場合はそれらは 1 個と数える）に対して、被験者に実験手順 2. と同じ評価基準で 1 位から 15 位の順位をつけてもらう。
4. 実験手順 3. で被験者によって 1 位から 15 位の順位が与えられた 15 個のキーワードを、それぞれを抽出した方法によって 5 種類 × 3 個に再分類し、各キーワード抽出法による 3 つのキーワードに与えられている順位を合計する。この合計順位の小さいキーワード抽出法から順に 1 位から 5 位の順位をつける。

表 5 に 2001 年 11 月における、実験に用いた各キーワ

表 4 注目キーワード (Hit 順)

Keyword	Note	New	Relation	Hit
テキスト	45	23	2.0	52
情報	44	24	1.9	45
システム	41	20	2.0	42
開発	21	15	1.4	34
環境	66	26	2.6	32
企業	64	22	2.9	32
人	26	24	1.1	31
研究	35	18	1.2	31
社会	35	21	1.7	27
ネットワーク	21	21	1.0	26
コンピュータ	28	28	1.0	25
活動	42	23	1.8	23
対応	44	22	2.0	21
総合	38	38	1.0	18
ツール	21	21	1.0	18

表 5 各キーワードと「創造」を共に含む Web ページ数の平均

	Note	New	Relat.	Freq.	Diff.
平均値	12.3	2.5	17.5	25.8	18.5

ードと「創造」という単語を共に含む Web ページ数の平均値を示す。頻度語 (Freq.) は提示した 15 個のキーワードと「創造」という単語を同時に含むページが平均で 25.8 ページと最も高い値を示しているのに対し、新規性 (New) による単語とでは最も低く、わずか 2.5 ページであった。また注目キーワード (Note) はほぼその中位にあたる 12.3 ページであり、高頻度語や低頻度語のどちらかに偏って選ばれていないことが分かる。表 6

*6 「創造」という語が出現する Web ページの内で、単語が出現している Web ページの数

に各キーワード抽出法に与えられた順位を示す。その結果、本稿で提案する注目キーワード(Note)と関連度によるキーワード(Relat.)とが共に高い評価を得た。これらのキーワードは「創造」という単語との関連が強く、その意味を理解しやすかったため創造との関係を容易に想像できたと考えられる。また、頻度(Freq.)による単語も「創造」との関連が強いと予想されるが、一般的であったり既知のありふれた単語が多くなったため評価が下がったと考えられる。新規性(New)によるキーワードが最も順位を下げている理由として、キーワードの出現頻度が低かったために、「創造」との関連を掴みにくかったことが挙げられる。すなわち、被験者への質問において「独創的な」という条件がついているため、周りと異なる頻度の低い語が選択される可能性もあったが、逆に独創的過ぎてもととの「創造」との関連が想像できずに、意味不明なノイズとなってしまったと考えられる。この点において、頻度が低い単語の出現する文脈を実験中に参照できなかったことは、新規性キーワードにはやや不利な条件であったと言える。しかしながら、表6において評価が最低の新規性による評価基準が、関連度と組み合わせられた注目キーワードにおいて、わずかではあるが関連度よりも評価が良くなっている。すなわち、多くのノイズを含む新規性が高い単語を、関連度によって有効な単語に絞り込むことでノイズを除去する効果が得られている。最近の伸びが高い単語(Diff.)に関しては、表5からもわかるように、一般に頻度の高い単語がWebページの増加に伴って、その頻度も増加しやすい傾向にあるため、単純に頻度が増加している単語が望ましいわけではないことが確認された。

本システムのユーザは出力される注目キーワードから有効なキーワードを選び出して自らの活動に適用する。表7に示すように被験者が選んだキーワード選択人数の分布は、キーワード抽出法によらずほぼ同数であり、2人以上に選ばれたキーワードは約6割にとどまっている。これは多くの被験者が自身の基準によってキーワードの取捨選択作業を行ない、出力として意味をなさなかったキーワードが取り除かれた結果である。すなわち、有効なキーワードはユーザによって異なり、その有用性も未知数であるが、不要なものを除く能力はみな共通に備えている。そのため新規性と関連度が共に高い注目キーワードをユーザが取捨選択する方が、それぞれを個別に取捨選択して取りまとめるよりも効率が良い。また、キーワード抽出法の組合せとして、関連語と頻度語(Freq.)や頻度の差(Diff.)によるキーワードとの組合せも考えられるが、頻度語は世の中で一般的にありふれた単語であり、将来の流行を

表6 キーワード抽出法の評価

順位	Note	New	Relat.	Freq.	Diff.
1位	8	3	7	3	2
2位	7	3	6	2	3
3位	1	3	2	5	9
4位	2	3	3	5	6
5位	2	8	2	5	0
順位合計	43	70	47	67	59

表7 15個のキーワードの選択人数の分布

人数	Note	New	Relat.	Freq.	Diff.
0-1人	7	6	4	5	8
2-6人	4	5	7	6	3
7人以上	4	4	4	4	4

予測するという目的に対しては頻度の低い単語の方が、現在はやっていないものを発見する手がかりとして有用であると考えられる。

5. 結論

本稿では、Web上のデータを用いて世の中の動向を確認しながら、流行を予測するための材料としての注目キーワードを出力するシステムを提案した。本システムはユーザの興味に関連しつつ、最近の世の中において出現機会が増えてきている語を注目キーワードとして出力し、これが多くのユーザの目に留まりやすいことを実験により確認した。人間が必要とするキーワードは興味がある単語との関連を想像でき、かつ一般的でない単語であった。また、人間の能力によって不要なキーワードは出力から除かれることを確認し、システムは少数の不要なキーワードを出力に含んでいても大きな問題にはならないことが確かめられた。注目キーワードは今後、ユーザの思考と世の中のデータとをつなぐ新たな観点として世の中のチャンスを発見するために用いられ、効果的なアクティブマイニングを実現するための創造活動[大澤 02]に役立つことが期待される。チャンスの気づきの主体は人間であり、我々の絶え間ない創造活動を助けるシステムとして、本稿で提案する手法を役立てたい。

参考文献

- [Beaulieu 97] Beaulieu, M.: "Experiments on Interfaces to Support Query Expansion", Journal of Documentation, Vol.53, No.1, pp.8-19, (1997).
- [Chasen] 松本裕治, 北内啓, 山下達雄, 平野義隆, 松田寛, 浅原正幸: 日本語形態素解析システム「茶筌」version2.0 使用説明書第二版, NAIST-IS-TR99012,

- (1999). (URL) <http://cl.aist-nara.ac.jp/lab/nlt/chasen/>
- [Fayyad 96] Fayyad, U., Piatetsky-S., G. and Smyth, P.: "From Data Mining to knowledge Discovery in Databases", *AI magazine*, Vol.17, No. 3, pp.37-54, (1996).
- [福田 01] 福田剛志, 森本康彦, 徳山豪: データマイニング, 共立出版, (2001).
- [Goldberg 87] Goldberg, D.E.: "Genetic Algorithms in Search", Addison-Wesley, (1989).
- [笠原 94] 笠原要, 松澤和光, 石川勉, 河岡司: 観点に基づく概念間の類似性判別, 情報処理学会論文誌, Vol. 35, No.3, pp.505-509, (1994).
- [小池 97] 小池雄一, 杉浦淳, 古関義幸: タイムマシン-時間情報に基づく WWW サイトの検索 I/F-, 第71回情報処理学会研究会報告, Vol.97, No.24, pp.37-44, (1997).
- [Luhn 58] Luhn, H.P.: The automatic creation of literature abstracts, In *IBM Journal for Research and Development*, Vol.2, No.2, pp.59-165, (1958).
- [大澤 02] 大澤幸生, 矢田勝俊, 庄司裕子, 鎌田真由美, 可合深雪: アクティブマイニング時代におけるチャンス発見の役割, 第47回人工知能学会基礎論研究会資料, pp.27-30, (2002).
- [Rocchio 71] Rocchio, J.: "Relevance Feedback in Information Retrieval", *The SMART Retrieval System-Experiments in Automatic Document Processing*, Prentice Hall Inc., pp.313-323, (1977).
- [Salton 97] Salton, G. and Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval, *Readings in Information Retrieval*, pp.323-328, (1997).
- [砂山 02] 砂山 渡, 谷内田正彦: 観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装, 人工知能学会誌, Vol.17, No.1, pp.14-22, (2002).
- [Yahoo] サーチエンジン Yahoo:
(URL) <http://www.yahoo.co.jp>.
(2002年12月2日 受付)
(2003年3月14日 採録)

[著者連絡先]

〒560-8531 大阪府豊中市待兼山町1-3
大阪大学大学院基礎工学研究科システム創成専攻
砂山 渡
TEL: 06-6850-6362
FAX: 06-6850-6341
E-mail: sunayama@sys.es.osaka-u.ac.jp

—— 著 者 紹 介 ——

すなやま わたる
砂山 渡 [非会員]

1995年大阪大学基礎工学部制御工学科卒業。1997年同大学院博士前期課程修了。1999年同大学院博士後期課程中退。同年同大学大学院助手、現在に至る。博士(工学)。人間の創造活動を支援する研究に興味をもつ。

やちだ まさひこ
谷内田 正彦 [非会員]

1971年大阪大学大学院工学研究科修士課程修了。同年同大基礎工学部助手、助教授、教授を経て1997年より同大学院基礎工学研究科教授、現在に至る。工学博士。画像処理、人工知能、移動ロボットなどの研究に従事。著書「ロボットビジョン」(昭晃堂)、「コンピュータビジョン」(丸善、編著)など。情報処理学会、ロボット学会など各会員。

Discovery of Notable Keywords from Web Pages for Estimating Future Trends

by

Wataru SUNAYAMA and Masahiko YACHIDA

Abstract :

Since there are many resources on WWW, it has been natural that we extract available information from them. In this paper, the system which extracts notable keywords is proposed. Notable keywords are related to user interests and newly appeared words in Web pages. That is, viewpoints and time are important factors. A subject of a topic varies from viewpoints, and the world changes time by time. Thus, notable keywords are useful for estimating future trends because Web pages represent real-time public opinions and user's viewpoint are reflected. Therefore, creative users are able to apply notable extracted keywords to their activities.

Keywords : Future trends, Notable keywords, WWW, Search Engines

Contact Address : **Wataru SUNAYAMA**

Dept. of Systems and Human Science,
Graduate School of Engineering Science, Osaka University
Machikaneyama-cho 1-3, Toyonaka, Osaka 560-8531, JAPAN
TEL : 06-6850-6362
FAX : 06-6850-6341
E-mail : sunayama@sys.es.osaka-u.ac.jp