具体性指向単語クラスタリングによる網羅的トピック発見と検索質問拡張支援

若木　裕美†　　正田　備也††　　高須　淳宏††　　安達　　淳††

† 東京大学大学院 情報理工学系研究科　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{hiromi,masada,takasu,adachi}@nii.ac.jp

あらまし　既存の検索エンジンではキーワード検索が主流で，自分の探したい分野について不慣れなユーザにとっては，数語からなる適切な検索質問を見つけるのが難しい．そこで，本研究では，ユーザによって入力された検索質問に関連する話題をトピックに分ける．また，個々のトピックを特徴付けるような単語群をユーザに提供し，ユーザがいずれかのトピックを選ぶことで，元々の検索質問の曖昧性を解消することを目的とする．このとき，ユーザが各単語群を見て新しい知識（固有名詞など）を発見できるような単語であることが望ましい．これらの目的に合致する単語を集めるために，本研究では，Tangibility をいう尺度を新しく定義した．既存の単語の重み付け手法では，トピックに関係なく一般的に使われる単語が多く得られるのに対し，Tangibility による重み付けでは，検索質問に含まれる様々なトピックを表す特徴的な単語が得られる．実験では，Tangibility と他の単語の重み付けの手法とを比較することでその性能を評価したので報告する．

キーワード　情報検索，データマイニング，クラスタリング，WWW

Exhaustive Topic Detection and Query Expansion Support Based on Substance-Oriented Term Clustering

Hiromi WAKAKI†, Tomonari MASADA††, Atsuhiro TAKASU††, and Jun ADACHI††

† Graduate School of Information Science and Technology, The University of Tokyo.
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan
†† National Institute of Informatics　2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †{hiromi,masada,takasu,adachi}@nii.ac.jp

Abstract　Conventional search engines are designed mainly for general keyword search. Users who are unfamiliar with the field they seek cannot find an appropriate combination of query terms. We aim at a query disambiguation method which discovers various topics buried in a search result by making clusters with the terms appearing in the retrieved documents. At that time, it is desirable that the terms should be key for users to discover new knowledge with term clusters, e.g. proper nouns. In this paper, we propose a new measure, called "Tangibility," to gather terms matching the above purpose. With existing term weighting methods, most of the high scored terms are generally used regardless of topics. On the other hand, with our Tangibility term weighting method, we can obtain distinctive terms each of which corresponds to a certain topic and all of which cover various topics implied by the query. We also conduct an experiment and compare our formula with the existing term weighting formulas.

Key words　Information Retreival, Data Mining, Clustering, WWW

1. はじめに

従来型の検索エンジンでは，ランキングされた結果を見ても所望の情報を得られるとは限らない．第一に，ランキングされた結果の中には，様々な粒度の情報や検索質問から類推される多義性を伴う内容が，一次元的に並べられてしまう．このため，ユーザは自分の要求に合致する内容を，何万もある検索結果の文書の中から探し出さなければならない．第二に，ユーザの入力する検索質問は，多くの場合 1，2 語であり，欲しい内容をそのまま示せるような特定の固有名詞などを含まない限り，検

索質問の中に曖昧性が存在しやすい．

つまり，結果の提示の仕方，大量にある結果の中からの絞込み，そしてユーザの検索質問自体の改善が，現在の検索エンジンに求められると考える．そこで本研究では，検索語をヒントに，より良い検索結果を導ける単語を集め，それをクラスタに分ける手法を提案する．適切なクラスタに分けることで，トピックのまとまりを表現することができ，また，多数を占めるトピックに押しつぶされてしまうトピックも拾い上げることができる．

すなわち,「検索語より具体性があり，検索語から連想するものとして適切であるが，検索語に包含される様々なトピックを網羅する」性質を持つ単語を集める．この性質を，Tangibility と呼ぶ．Tangibility のある単語が適切に選ばれていれば，その単語を用いて適切にクラスタリングするだけで，多様なトピックを発見することができる．このように，Tangibility を持つものとして選ばれた単語を幾つかのトピックに分けて提示することで，ユーザは自分の検索要求の中にある曖昧性に気が付き，また，提示された単語を利用して自分の要求に合致する検索結果を容易に集めることが可能になると考える．実験では，Tangibility と他の単語の重み付けの手法とを比較することでその性能を評価した．

2. 先 行 研 究

キーワード 抽出（重要語抽出）の手法は，既存の研究でも数多く提案されている．その中でも，語の分布を測り代表性 (representativeness) という指標にあう語を重要とみる手法 [8] や，語の共起に χ^2 検定を利用して重要な語を選ぶ手法 [12] が，語の分布のずれを見るという観点では近い．しかし，全文書における単語の重要度（あるいは各文書中での単語の重要度）で一次元的に並べること考えており，その点では今回の目的と異なる．

今後は，ユーザの入力した語を手がかりに，その曖昧性を解消するため，あるいは，より詳細な情報を入力するために，システムが情報を提供することが必要となると考えている．そこで，検索語を元に Web 上にあるデータをトピックに分け，曖昧性を指摘できれば，そのいずれであるかをユーザは答えることが可能となるであろう．

クラスタの可視化という点では，語の共起を可視化してキーワードを抽出する手法：KeyGraph [6] [14] や，対話性を重視した検索インターフェースを持ち，特徴語グラフを表示できる DualNAVI [7] といった研究が行われている．しかし一般のユーザに提示する場合には，一見して内容が分かるように表示する必要があると考える．そこで，幾つかの単語をクラスタとして表現することで，ユーザが利用しやすいように提示することを目的とした．

3. Tangibility について

3.1 定　　義

本研究では，単語が同じ 文書の中で同時に出現することを，単語の共起と言う．単語 t_i と単語 t_j が共起する回数は，単語 t_i と単語 t_j が同時に出現した文書の数によって定義する．また，単語 t_i の出現確率を $P(t_i)$ と書き，単語 t_i が出現する文書数を全文書数で割った値と定義する．単語 t_i が出現する文書において単語 t_j が出現する確率を $P(t_j|t_i)$ と書き，単語 t_i と t_j が共起する文書数を単語 t_i が出現する文書数で割った値と定義する．単語 t_i の document frequency(以下，DF と呼ぶ) を $DF(t_i)$ と書くことにする．

3.2 Tangibility の仮説

本研究は，ユーザが自分の検索質問を改善するために用いることのできる語群の発見を目的とする．そのためには，最初の検索語によって得られた検索結果の中から得ることができ，かつ，検索結果に含まれる多様なトピックを弁別するために有用な単語を見つけ出すことが必要となる．このような性質を，Tangibility と呼ぶことにする．そこで，本研究では，このような単語は「特定の語群とのみよく共起する単語である」という仮説を立てた．そして，この仮説を実験によって検証することにした．

Tangibility をもつ単語に期待されることは，検索語より具体性があり，検索語から連想するものとして適切であるが，検索語に包含される様々なトピックを網羅することである．そこで，Tangibility をもつ単語を選ぶための単語への重み付けとして，本研究では下記のような定式化を提案し，これを TNG3 と呼ぶ．

まず，単語 t_j の出現頻度が，単語 t_i が存在するという状況が加わることによって，どれだけ増大するかを，次の値によって評価する．

$$\Delta_{t_i}(t_j) \;=\; P(t_j|t_i) \times \log \frac{P(t_j|t_i)}{P(t_j)} \tag{1}$$

ここで，$F_i = \{t_j | \Delta_{t_i}(t_j) > 0\}$ とし，下記のように Tangibility の定式化 TNG3 を得る．

$$TNG3(t_i) = \frac{\sum_{t_k \in F_i} (\Delta_{t_i}(t_k) \times DF(t_k))}{|F_i|} \tag{2}$$

「特定の語群とのみよく共起する」ことの定式化は，筆者らが他の文献でも行っており，検索における性能向上は確認済みである [10] [11]．ただし，[10] [11] での定式化では，一定の方法で選び出された文書の集合と，それらの文書が選び出された元の文書集合という，二つの文書集合を必要とした．しかし，本稿の定式化では，以前の定式化をより洗練させ，ある検索質問によって得られた検索結果など，一つの文書集合のみから，同様の効果を得られるように改善した．したがって，そのような文書集合が選び出される元の文書集合を必要としないため，より多様な文書集合に用いることが可能である．

3.3 Tangibility の式の意味

情報量のひとつに KL 情報量という量がある．語の共起に関連して意味を考えると,『単語 t_i の**出現**が，別の単語 t_j の**存在**に，どれだけ影響するか.』ということを表す量である．このとき，KL 情報量は次の式で表される．

$$\begin{aligned} KL(t_j;t_i) = \quad & P(t_j|t_i) \log \tfrac{P(t_j|t_i)}{P(t_j)} \\ & + P(\neg t_j|t_i) \log \tfrac{P(\neg t_j|t_i)}{P(\neg t_j)} \end{aligned} \tag{3}$$

TNG3 のための式 (1) は，式 (3) の前半部分の項 $P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)}$ と等しい．つまり，式 (1) は,「単語 t_j が出現する確率に比べて，単語 t_i が出現するときに単語 t_j が出現する確率が増えるかどうか」を測るものである．増える場合には，この項は 0 より大となる．

TNG3 では，$F_i = \{t_j|\Delta_{t_i}(t_j) > 0\}$ という条件を満たす場合のみ，式 (1) が式 (2) の中で用いられる．特に，$|F_i|$，すなわち，$\Delta_{t_i}(t_j)$ が 0 より大となった単語 t_j の個数で割ることで，単に増大した量の和ではなく，単語ひとつ当りの平均でどれくらい大きく増大したかを算出する．もし，$\Delta_{t_i}(t_j)$ が 0 より大となる単語 t_j の個数で割らずに，総和をそのまま TNG3 の値とすると，次のような不都合が生じる．つまり，単語 t_i が出現しているという条件をつけることで，多くの単語の出現頻度が少しずつ増えるという状況と，特定の単語だけその出現頻度が大きく増えるという状況（単語 t_i が Tangibility を持つ状況）とを，区別できないこととなる．このように,「特定の単語と**のみ**，**よく**共起する」という Tangibility の仮説を忠実に定式化したのが式 (1)，式 (2) である．

3.4 単語間の類似度

単語のクラスタリングにおいて，単語間の類似度をどのように定義するかは，重要な要素のひとつである．本研究では，単語 t_i と t_j の類似度を，次のように定義する．

$$Sim(t_i,t_j) = \frac{\text{単語 } t_i \text{ と } t_j \text{ が共起する文書数}}{\text{単語 } t_i \text{ と } t_j \text{ の少なくとも一方が現れる文書数}} \quad (4)$$

3.5 クラスタリングのアルゴリズム

まず，3.4 で定義した類似度が近い順に，単語のペアを列挙していく．そして，各単語のペアについて，次の条件を用いて，クラスタを形成する要素とするかどうかを判断する．

（1） 単語 t_i と t_j の両方が，すでに列挙されたペアに含まれないならば，t_i と t_j だけを含む新しいクラスタを作る．

（2） 単語 t_i と t_j のどちらかが，すでに列挙されたペアに含まれるとき，単語 t_i が含まれるクラスタを C_1，単語 t_j が含まれるクラスタを C_2 とする．$s(C_1,C_2)$ をクラスタ C_1 と C_2 間の重み，$s(C_1,C_1)$ をクラスタ C_1 内の重みとする．クラスタ内の重みは，クラスタ内に存在する単語のすべての組の類似度の総和，クラスタ間の重みは，対象とするふたつのクラスタのそれぞれに属する単語同士のすべての組の類似度の総和であり，次のように定式化する．

$$s(C_1,C_2) = \sum_{t_i\in C_1}\sum_{t_j\in C_2} Sim(t_i,t_j)$$

$$s(C_1,C_1) = \sum_{t_i\in C_1}\sum_{t_j\in C_1} Sim(t_i,t_j)$$

そして，

$$\frac{s(C_1,C_2)}{(s(C_1,C_1))\times(s(C_2,C_2))} \geqq \tau \quad (5)$$

を満たせば，クラスタ C_1 と C_2 を結合する．また，τ を本稿におけるクラスタリングパラメータとする．

ただし，単語 1 個の場合もクラスタとみなす．例えば，単語 t_i だけが既出で，単語 t_j が新出の場合には，t_j は単語 1 個でクラスタを形成するものとみなし，このクラスタを t_i が属するクラスタと結合して良いかどうかを判定する．その際，t_j のみを含むクラスタ内の重みは，上の定義より $Sim(t_j,t_j)$ となり，この値は 1 に等しい．

クラスタリング・アルゴリズムには，階層的クラスタリング，分割的クラスタリング，確率的クラスタリング，グラフ理論的クラスタリングなど，様々な種類がある [4]．だが，本稿の目的はクラスタリング手法を比較することではないので，クラスタの結合判定については，Ding ら [2] の研究を参考にし，上記のように平易な方法を採用した．しかし，クラスタリング・アルゴリズムについては，Ding らの提案しているアルゴリズムが，クラスタを順次結合することによる階層的クラスタリングを，そのつどクラスタのあらゆるペアの類似度を評価することで厳密におこなっているため，単語数が増えるにつれて計算時間が急激に増加してしまう．そこで，本研究では，クラスタのあらゆるペアの類似度を評価するのではなく，単語のペアを類似度の高い順に付け加え，そのペアが関係する二つのクラスタの類似度を計算するにとどめ，現実的な計算時間を得た．また，実験の結果は，適切な単語が選択されていれば，このように簡素化されたクラスタリング・アルゴリズムによっても，性質のよい単語クラスタが得られることを示している．

4. 実　　験

4.1 実験における比較対象

今回の実験において，Tangibility による式 (2) と比較する式は，単語の重み付けによく用いられる相互情報量 (Mutual Information: MI) [9] を含む式，そして，検索質問拡張の際に用いられる Robertson's Selection Value(RSV) [5] である．これら三種類の式によって単語に重みを与え，その重みが大きい順に一定数の単語を選び，クラスタリング・アルゴリズムの入力とする．また，document frequency(DF) もしばしば単語の重要性を示す指標として用いられる [13] ため，DF が大きい順に一定数の単語を取り出したケース (DFL) と，逆に DF が少ない順に単語を取り出したケース (DFS) も比較対象とした．ただし，DFS では純粋に小さい方から取ると DF が 1 の単語だけが取り出されてしまう．そして，DF が 1，つまり，一つの文書にしか現われない単語は，ほとんどの場合，与えられた文書集合に含まれる様々なトピックを表す単語としては，不適切である．そのため，DFS のケースにおいては，DF が 10 以上のものに限定して，DF が小さい順に単語を取り出すことにした．

単語の重みを定める式の基本的な形は，

$$W(t_i) = \frac{\sum_{t_j}(w(t_i,t_j)\times DF(t_j))}{|\{t_j|w(t_i,t_j)>0\}|} \quad (6)$$

$w(t_i,t_j)$ = 単語の共起情報から求められる，単語 t_i と t_j の関連度を示す値

であり，$w(t_i,t_j)$ の部分を各重み付けの式 (TNG3，MI) に置き換える．RSV では，単語の共起情報を使っていないため，RSV

の式をそのまま使って単語の重み付けを行う．DFL，DFS では DF の大きさに従って，単語を取り出す．

4.1.1 MI

直感的には，単語 t_i と t_j に関する相互情報量は，単語 t_i（あるいは単語 t_j）の有無を知ることで単語 t_j（あるいは単語 t_i）の有無について得られる知識の量を示す量であり，次のような式で書ける．

$$\begin{aligned} MI(t_i,t_j) = \quad & P(t_i)\{P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} \\ & +P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)}\} \\ & +P(\neg t_i)\{P(t_j|\neg t_i)\log\frac{P(t_j|\neg t_i)}{P(t_j)} \\ & +P(\neg t_j|\neg t_i)\log\frac{P(\neg t_j|\neg t_i)}{P(\neg t_j)}\} \end{aligned}$$

この MI を二つの単語の関連度を示す値として用いると，単語への重み付けの式は下記のようになる．

$$W(t_i) = \frac{\sum_j (MI(t_i,t_j)\times DF(t_j))}{|\{t_j|w(t_i,t_j)>0\}|} \tag{7}$$

4.1.2 RSV

RSV (Robertson's Selection Value) [5] は，検索質問拡張に使われる単語の選択のための特徴量であり，特定の検索語による検索結果など，一定の仕方で選ばれた文書集合に対して各単語がどの程度強く関連しているかを評価するために使われる．RSV は以下の式で定義される．

$$RSV = (\frac{rdf}{R}-\frac{df}{N}) \times\left\{\alpha\times\log\frac{N}{df}+(1-\alpha)\times\log\frac{\frac{rdf+0.5}{R-rdf+0.5}}{\frac{df-rdf+0.5}{N-df-R+rdf+0.5}}\right\} \tag{8}$$

- rdf : 検索結果など，一定の方法で選択された文書の集合 S の中で，単語 t を含む文書数
- R : 文書集合 S に含まれる文書数
- df : 文書集合 S がそこから取ってこられた文書集合全体 U の中で，単語 t を含む文書数
- N : 文書集合 U に含まれる総文書数
- α : パラメータ

今回の実験では，α を 0.5 に設定した．なお，この特徴量は単語の共起情報を利用しない．しかし，単語の現われ方の偏りを調べるために，基準となる単語の出現頻度を必要とするため，十分に大きな文書の集合（上に U と示した集合）が与えられていなければならない．

4.2 実験の概要

4.2.1 用いたデータ

実験には，Web 上にある産経スポーツ(注1)のバックナンバーを利用した．サッカー，日本の野球，メジャーリーグの三つの分野のいずれかに属する文書を集め，これを実験対象とした．文書の総数は 3519 個であった．

特に RSV では，部分的な文書集合以外に，これを包含する全体の文書集合が必要になるため，NTCIR3 および NTCIR4 のために準備されたコーパスである NW100G-01 [3] を仮想的な全体集合とみなし，そこでの単語の出現頻度を基準として，単語の現われ方の偏りを求めることにした．この NW100G-01 には約 1000 万件の文書が含まれている．RSV の式 (8) 中の R は，産経スポーツから得た文書数が 3519 個であることから 3519 とし，N は，NW100G-01 に含まれる文書の総数 (10253810 個) と産経スポーツから得た文書数の合計 10253810＋3519 とした．

4.2.2 実験の方法

（1） 単語の重み付けの各手法により得られた順位のうち，上位 400 語の単語を集める．

（2） それらの全ての組について距離を測り，距離の近い順に上位 8000 番目までの組についてクラスタリングを行う．

（3） 得られた単語のクラスタについて，適合率を測る．

4.2.3 評価の方法

本稿では，単語のクラスタリングを実験目的とし，その評価を行う．しかし，筆者等の知る限り，与えられた文書集合に現れる単語の集合を，その文書集合に含まれるひとつひとつのトピックに対応するように分けるクラスタリングについて，その良し悪しを評価するための標準的な方法は，これまで提案されていない．そこで代わりに，正解としての分類データが付いている文書の集合を利用して，次のように評価を行う．

本実験では，Web 上の産経スポーツのバックナンバーから，日本の野球，メジャーリーグ，サッカーの 3 つのトピックに属する文書を集め，実験対象とした．そこで，D_1，D_2，D_3 を，それぞれ日本の野球，メジャーリーグ，サッカーに対応する文書の集合とする．

また，これら 3 つの文書集合における単語 t の DF の値を，それぞれ $DF(t;D_1), DF(t;D_2), DF(t;D_3)$ と表わす．$DF(t)$ は，文書集合全体 $D_1\cup D_2\cup D_3$ における単語 t の DF を表わすとする．このとき，ある単語クラスタ C に対応するトピックを，$\sum_{t\in C} DF(t;D_k)$ を最大にする文書集合 D_k によって定める．例えば，単語クラスタ C について，$\sum_{t\in C} DF(t;D_1)$ と $\sum_{t\in C} DF(t;D_2)$ と $\sum_{t\in C} DF(t;D_3)$ の三つの値のうち，$\sum_{t\in C} DF(t;D_2)$ が最大なら，C に対応するトピックはメジャーリーグである．そして，単語クラスタ C の精度 $Prec(C)$ を，下記の式によって定義する．

$$Prec(C) = \frac{\sum_{t\in C} DF(t;D(C))}{\sum_{t\in C} DF(t)}$$

ただし，$D(C)$ は，単語クラスタ C に対応するトピックの文書集合を意味する．上の定義より，単語クラスタ C の精度は，クラスタに含まれるすべての単語が，同じ一つのトピックの文書集合に偏って出現しているとき，高くなる．

次に，各クラスタの精度をもとにして，クラスタリング結果全体の精度を求める方法について議論する．各クラスタの精度を総合する方法には，macroaveraged precision と microaveraged precision という二通りがある [1]．macroaveraged precision は，

(注1)：http://www.sanspo.com/

$$\frac{\sum_{\text{全クラスタ}} \text{クラスタの精度}}{\text{クラスタの個数}}$$

と定義され，microaveraged precision は，

$$\frac{\sum_{\text{全クラスタ}} \text{クラスタに含まれる正解の個数}}{\sum_{\text{全クラスタ}} \text{クラスタに含まれる要素の個数}}$$

と定義される．

本実験においては，microaveraged precision を評価尺度として用いた．なぜなら，macroaveraged precision では，個々のクラスタのサイズに関係なく，すべてのクラスタの精度の単純平均をとるため，クラスタのサイズが小さいほど精度が良くなりやすい問題設定の下では，小さいクラスタサイズが数多く生成されるほど，クラスタリング結果全体の精度が高いと判定されやすい．一方，microaveraged precision では，各クラスタの精度にクラスタのサイズを乗じてから和をとるため，大きなクラスタの適合率が，クラスタリング結果全体の評価に効きやすい．すなわち，精度の高い小さなクラスタが沢山できるよりは，精度の高い大きなクラスタが得られているほうが，評価上有利となる．今回の実験で用いる文書集合は，三つのトピックを含んでいるだけであるから，なるべく大きく，かつ正確な単語のクラスタが構成されているかどうかが問題となる．また，Tangibility をもつものとして選択された単語は，そもそも一つのトピックの文書に偏って出現しやすいため，macroaveraged precision による評価では，すべての単語を孤立させる自明なクラスタリングの評価が自然に高くなってしまう．そこで，microaveraged precision を評価方法として採用した．

評価したい単語クラスタの集合を $\mathcal{C}$ とし，クラスタリング結果全体の精度 $Prec(\mathcal{C})$ を，次のように定義する．

$$Prec(\mathcal{C}) = \frac{\sum_{C\in\mathcal{C}} \sum_{t\in C} DF(t; D(C))}{\sum_{C\in\mathcal{C}} \sum_{t\in C} DF(t)}$$

この定義では，単語クラスタのサイズとして，クラスタに含まれる単語の個数ではなく，クラスタに含まれる単語の DF の総和を用いている．なぜなら，クラスタごとの精度の定義での分母が，クラスタに含まれる単語の DF の総和であるため，これをクラスタのサイズと解釈するのが自然だと考えたからである．

ここで，クラスタリング結果の評価の例を示す．得られたクラスタが表 1〜3 の三つのクラスタ C_1, C_2, C_3 であるとする．まず，各クラスタに対応するトピックが，日本の野球，メジャーリーグ，サッカーとそれぞれ判定される．次に，C_1 では (表 1),

$$\sum_{t\in C_1} DF(t; D(C_1)) = 1303 + 1312 + 1222 = 3837$$

$$\sum_{t\in C_1} DF(t) = 3911$$

C_2 では (表 2),

$$\sum_{t\in C_2} DF(t; D(C_2)) = 850 + 1017 = 1867$$

$$\sum_{t\in C_2} DF(t) = 2022$$

C_3 では (表 3),

$$\sum_{t\in C_3} DF(t; D(C_3)) = 581 + 381 + 283 + 526 = 1771$$

$$\sum_{t\in C_3} DF(t) = 2401$$

であることから，

$$Prec(\mathcal{C}) = \frac{3837 + 1867 + 1771}{3911 + 2022 + 2401} = \frac{7475}{8334} = 0.8969\ldots$$

のように，クラスタリング結果全体の精度が求まる．

表 1　例．クラスタ C_1

単語	D_1 での DF	D_2 での DF	D_3 での DF	計
楽天	1303	1	0	1304
巨人	1312	59	3	1374
ロッテ	1222	7	4	1233
計	3837	67	7	3911

表 2　例．クラスタ C_2

単語	D_1 での DF	D_2 での DF	D_3 での DF	計
中田	138	0	850	988
海外	17	0	1017	1034
計	155	0	1867	2022

表 3　例．クラスタ C_2

単語	D_1 での DF	D_2 での DF	D_3 での DF	計
松井	98	581	295	974
ゴジラ	66	381	0	447
メッツ	68	283	0	306
イチロー	102	526	1	629
計	335	1771	295	2401

4.3　結果と考察

4.3.1　クラスタリング精度の比較

図 1 は，クラスタの粒度と，クラスタリングの microaveraged precison $Prec(\mathcal{C})$ の関係を示している．横軸が，クラスタリングのパラメータ τ であり，この値が大きいほどクラスタは小さくなる．特に，値が 1 のときは，選ばれた単語の多くが単独でクラスタをなしている状態 (粒度が細かい状態) であり，このときクラスタリング結果の精度は，選ばれた単語のひとつひとつが特定のトピックの文書集合にどれだけ偏っているかを近似している．図 1 によると，パラメータが 1 のときには，TNG3, DFS, MI, RSV, DFL の順で精度が高い．つまり，TNG3 によって選ばれた単語群が，特定のトピックへ関連する傾向を最も強く示していると言える．また，粒度の粗い側を見ると，パラメータが 0.001 になる前のところで，全ての手法において精度が大きく下がっている．つまり，手法によらず 0.001 以下ではパラメータの設定が不適切であることを示す．実際のクラスタリング結果を調べてみると，パラメータが 0.001 にまで小さ

くなると，どの手法でも，日本の野球，海外の野球，サッカーに関する単語が，すべて同じクラスタに入ってしまっていた．

中間的な粒度の場合を見ると，特にパラメータが 0.005 付近では，TNG3 だけが高い精度を保っている．これは，TNG3 によって選ばれた単語が，単に個々に特定のトピックに強く関係しているだけでなく，クラスタリングによってまとめられた後でも，出現の偏りが保たれるような性質を持っていることを示している．このことは，パラメータが 1 のときの精度が二番目に良かった DFS と比較すると，分かりやすい．つまり，DFS によって選ばれた単語は，そのひとつひとつを見れば特定のトピックに強く関係してはいる．しかし，DFS では，粒度を粗くするにしたがって，精度が着実に低下する．これは，DFS によって選ばれた単語のペアのうち，それらが共起する文書の数が多いもの（つまりクラスタリングの過程で結合されやすい単語のペア）については，その文書群が複数のトピックに分散しやすいためだと考えられる．一方，TNG3 の場合には，パラメータが小さい範囲まで安定した精度を維持している．これは，TNG3 によって選ばれた単語のペアをとると，それらが共起する文書の数が多くても，その文書群が一つのトピックにかたまりやすいためだと考えられる．ここに，本研究の提案する Tangibility の特徴が現れていると言える．つまり，単に DF が小さめの単語を集めるだけでは，文書集合に含まれるトピックの各々に対応する単語のグループをうまく得られない，ということである．実際に，TNG3 で作られたクラスタに含まれる単語群の例を，表 5，表 6 に載せた．

4.3.2 クラスタの単語数とクラスタ毎の精度の関係の比較

図 2,3,4 は，クラスタに含まれる単語の数と，クラスタ毎の精度 $Prec(C)$ の関係を，TNG3，MI，RSV で得られた単語のクラスタリング結果について，それぞれ示している．また，各々の図は，クラスタリングのパラメータを 0.05，0.005，0.001 にした 3 通りの結果を含んでいる．パラメータが 0.05 の場合，TNG3，MI，RSV のいずれにおいても目立った大きなクラスタは無い．TNG3 はクラスタあたり 1〜30 個程度の単語を含み，MI と RSV は 1〜40 個程度の単語を含んでいる．しかし，精度を見ると，TNG3 では 0.8 以上の部分に多くのクラスタが分布しているのに対し，MI や RSV では 0.4 程度の低いところまで広く分布している．これが，図 1 において，MI や RSV の場合で全体的に精度を下げる結果となっていることが理解できる．クラスタリング・パラメータが 0.005 や 0.001 になると，TNG3 ではクラスタの単語数が大きくなっても精度が下降しない．このことから，TNG3 では適切に単語がクラスタリングされたといえる．それに対し，MI と RSV では 1 つのクラスタに単語が集中してしまい，最大のクラスタに含まれる単語数は 150 を超える．このとき，適切に単語が集められていないために適合率は下がっている．

また，いずれのパラメータをとってみても，TNG3 では単語数 1 のクラスタはほとんどないのに対し，MI と RSV では単語数 1 のクラスタが多く存在する．実際に結果のデータを見てみると，スポーツ全般に関係するが DF は小さい単語が，1 つの単語のみのクラスタとなってしまっている．つまり，MI や RSV では，特定のトピックにしか関係しない単語ではなく，スポーツ全般に関係するような単語を，DF が比較的小さい単語の中から拾ってしまっている．一方，TNG3 では，このような単語が選ばれにくい．なぜなら，TNG3 では，DF の大小にかかわらず，似たような文書群に出現する単語が他にもいくつか存在するときに限り，単語に高い重みが与えられるからである．「自分が出現する文書群と似たような文書群に出現する単語が，他にもいくつか存在するかどうか」という観点は，MI や RSV にはない観点である．

クラスタリングの対象となる単語は，各手法によって選ばれた上位 400 語であるが，実際にクラスタリングの過程でチェックされる単語は，400 語から作られる $400(400-1)/2$ 個のペアのうち，距離の近い順で上位の 8000 ペアのみである．今回のアルゴリズムでは，このペアの中に出てこない単語があれば，その単語ひとつだけでクラスタを構成するとみなされる．したがって，類似度が高い単語のペアに入るような単語を選択できなければ，クラスタリングがうまくいかないようになっている．この意味でも，MI や RSV で選ばれた単語より，TNG3 によって選ばれた単語は，効果的に機能しているといえる．

4.3.3 選択された単語の DF の分布の比較

図 1 において，精度が TNG3 に最も近いのは，DFS である．そのため，TNG3 によって選ばれた単語の弁別性が高いのは，TNG3 が DF の小さいものばかりを集めているためであると思われるかもしれない．そこで，表 4 において各手法で得られた単語の DF の分布を比較してみる．まず，今回用いた産経スポーツの記事は全部で 3519 文書であり，そこに含まれるトピックは 3 つであるため，全文書数の 3 分の 1 を超えて出現する単語は，色々な単語と共起してしまい，各トピックに対応するように分かれたクラスタを構成するのには有用でないと思われる．TNG3 では，DF が 1000 以上の単語は比較的少なくなっており，複数のトピックに関係する単語を含む曖昧なクラスタを生み出しやすい単語は，選ばれにくいことが分かる．その反面，DF が 100 未満の範囲では，MI や RSV では単語の個数が 0 となっているのに対し，TNG3 では 164 個の単語が分布している．しかし，DF が小さいという理由だけによって，TNG3 によって選ばれた単語が，クラスタリングの精度を向上させているわけではない．実際，DFS の場合には，クラスタリングの精度が TNG3 の場合より低かったにもかかわらず，400 個の単語すべての DF が 20 以下であり，DFS に比べると，TNG3 では，比較的 DF が大きな単語も，バランスよく選ばれている．

5. おわりに

本稿では，単語の具体性を数値的に測ることができる Tangibility という指標を提案した．また，Tangibility の定式化である TNG3 の式を提案した．TNG3 は情報量的な式からの改善であり，Tangibility の仮説を適切に定式化できたと考えている．更に，比較対象とした MI，RSV，DFS，DFL の各手法に比べて，TNG3 の適合率は非常に高く，クラスタリングへの寄与が大きいことが分かる．クラスタサイズや各単語の持つ DF についても調査し，TNG3 の式の有効性を確認した．

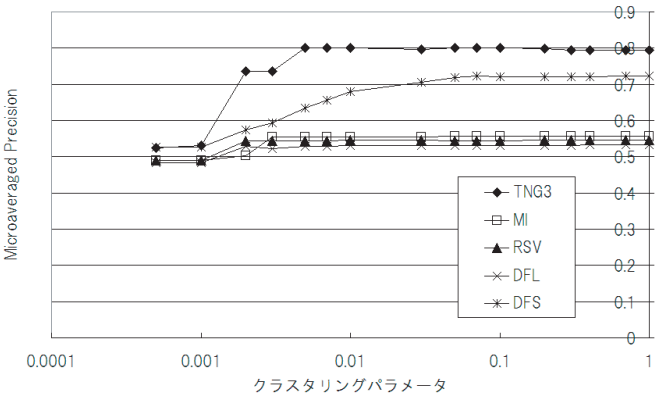

図 1　各手法について，クラスタリングパラメータを変化させたときの Microaveraged Precision の比較.

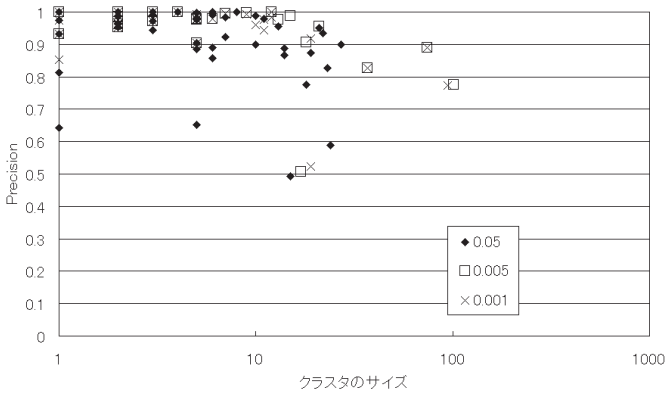

図 2　TNG3 の Microaveraged Precision の値と，クラスタのサイズ．クラスタリングパラメータは 0.05，0.005，0.001 の 3 種類を比較.

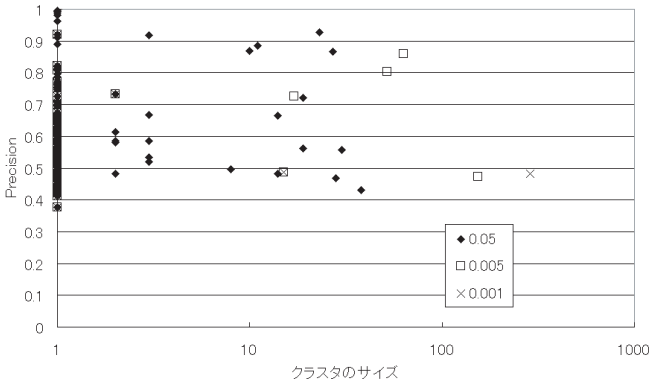

図 3　MI の Microaveraged Precision の値と，クラスタのサイズ．クラスタリングパラメータは 0.05，0.005，0.001 の 3 種類を比較.

本手法は，検索質問に含まれる曖昧性（多義性）を，それぞれのトピックに分けてユーザに提示できるシステムを最終的な目標としている．Tangibility のある単語 1 語を追加する場合

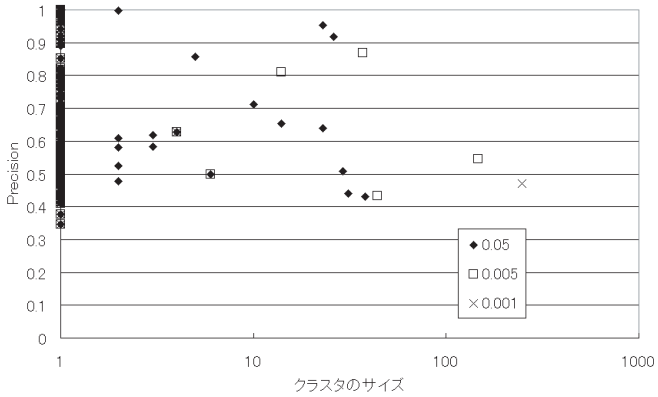

図 4　RSV の Microaveraged Precision の値と，クラスタのサイズ．クラスタリングパラメータは 0.05，0.005，0.001 の 3 種類を比較.

表 4　TNG3，MI，RSV，DFS の各手法によって得られた単語 400 語についての，DF の分布（単位：単語数）

DF	TNG3	MI	RSV	DFS
1500 以上	7	72	91	0
1000〜1500	17	34	27	0
500〜1000	64	159	110	0
250〜500	87	135	144	0
100〜250	61	0	28	0
20〜100	60	0	0	0
10〜20	94	0	0	400
10 未満	10	0	0	0

の検索質問拡張の性能は，筆者らの [10] [11](注2)で確認済みであり，今後はクラスタリングした単語群を検索拡張に用いる実験を行う予定である．

文　献

[1] Soumen Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

[2] C. Ding and X. He. Cluster merging and splitting in hierarchical clustering algorithms. In *IEEE International Conference on Data Mining (ICDM'02)*, pp. 139–146, 2002.

[3] Oyama K. Ishida E. Kando N. Eguchi, K. and K. Kuriyama. Overview of the web retrieval task at the third ntcir workshop, 2003.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264–323, 1999.

[5] Toyoda Masashi, Kitsuregawa Masaru, Mano Hiroko, Itoh Hideo, and Ogawa Yasushi. University of tokyo/ricoh at ntcir-3 web retrieval task. In *Proc. of the 3rd NTCIR Workshop Meeting*, pp. 31–38, 2002.

[6] Y. Ohsawa, E.B. Nels, , and M. Yachida. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proc. of IEEE ADL'98*, 1998.

[7] A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi, and H. Sakurai. Associative information access using dualnavi. In *Proc. of ICDL'00*, pp.

(注2)：[10] [11] 中では，Articulateness（分節性）と呼んでいるが，本稿にて Tangibility と改名した．また，同様に定式化として AR1，AR2 を定義しているが，本稿ではそれに続く新しい式として，TNG3 を定式化した.

表 5　TNG3 で得られた単語について，クラスタリングパラメータが 0.05 のとき，サイズの大きな上位 5 つのクラスタ.

クラスタサイズ	$D(C)$	クラスタ自身の適合率	単語群
27	サッカー	0.898626	中田 英 海外 Ｊリーグ 俊輔 ＦＷ ＭＦ 代表 高原 W杯 ドイツ 柳沢 予選 杯 欧州 ゴール メッシーナ ＤＦ ハンブルガー ジーコ 鹿島 東京 大阪 移籍 小野 アウエー ホーム
24	日本の野球	0.589595	勝 敗 目 安打 回 失点 投手 勝利 連続 打 差 手 連敗 連勝 回戦 打点 番 本塁打 負け サヨナラ 登板 号 弾 球
23	メジャーリーグ	0.827410	田口 井口 イチロー 稼 松井 秀 央 ソックス レッド ゴジラ 共同 軍 ヤンキース 大塚 ヤ インディアン ドレス メッツ ホ 地区 ラスベガス 高津 打数
22	日本の野球	0.933853	楽天 巨人 ロッテ オリックス 西武 広島 中日 大学 ヤクルト 高校 日本ハム 球団 横浜 鷹 古田 プロ ソフトバンク 野村 阪神 来季 青木 燕
21	日本の野球	0.950072	駒大 苫小牧 虎 竜 アマ 投 打線 ＫＯ 星野 対抗 借金 仰木 完投 田尾 山本 甲子園 完封 渡辺 西口 新井 ケタ

表 6　TNG3 で得られた単語について，クラスタリングパラメータが 0.005 のとき，サイズの大きな上位 5 つのクラスタ.

クラスタサイズ	$D(C)$	クラスタ自身の適合率	単語群
101	日本の野球	0.776698	楽天 巨人 ロッテ オリックス 西武 広島 中日 大学 ヤクルト 高校 日本ハム 球団 横浜 勝 敗 目 安打 回 失点 投手 鷹 古田 勝利 連続 打 差 プロ 手 ソフトバンク 連敗 野村 連勝 阪神 回戦 来季 青木 キャンプ 秋季 燕 打点 番 駒大 苫小牧 本塁打 ＦＡ 交渉 野口 行使 城島 残留 ＫＯ アマ 負け サヨナラ 登板 号 打線 氏 投 契約 弾 ブラウン 球 ボビー 仰木 甲子園 三振 高校生 巡 ドラフト 虎 竜 清原 谷 打席 星野 対抗 松坂 落合 借金 完封 山本 セ 辻 完投 メジャー 田尾 渡辺 牛島 会談 自己 ケタ 西口 斉藤 岩隈 新井 被弾 青学大 東都 新庄 ＷＢＣ
74	サッカー	0.890372	中田 英 海外 Ｊリーグ 俊輔 ＦＷ ＭＦ 代表 高原 W杯 ドイツ 柳沢 予選 杯 欧州 ゴール メッシーナ ＤＦ ハンブルガー ジーコ ン ボルト プレミア イングランド カズ 鹿島 東京 大阪 移籍 セリエＡ 小野 組 平山 ヘラクレス アウエー セルティック ホーム クラブ 豪州 田中 達 浦和 後半 柏 ラモス 協会 デビュー 招集 小笠原 フランス ジダン フル ナウ ジーニョ マド リード レアル スペイン ブラジル ベッカム 負傷 全治 ルマン 中沢 イラン アシスト 韓国 節 オランダ 大黒 巻 ヒデ 欠場 マルセイユ バイエルン
37	メジャーリーグ	0.828181	バファロー 田野 田口 井口 イチロー 稼 松井 秀 央 ソックス レッド ゴジラ 共同 軍 ヤンキース 大塚 ヤ インディアン ドレス メッツ 木田 タコマ ホ 地区 ラスベガス カージナルス 壮 高津 打数 アスレチック 藪 マルチ ボンズ 野茂 石井 ノーフォーク 大家
21	サッカー	0.954717	ラゾビッチ 在席 伯 時間切れ ボンフレール 歯医者 専守防衛 虫歯 脳しんとう バロシュ バドン 店長 セルティックショップ 釜本 万感 ジェラード サントス リバプール スシボンバー 塩田 チェルシー
18	日本の野球	0.906100	アンパン スチュワート 取り掛かる 毎度 山村 ロマノ 陽東 宮越 パスクチ 学院 青森 柳田 ツバメ 乱調 鳴門 ダルビッシュ 全日本 激怒

285–289.

[8] Hisamitsu Toru, Niwa Yoshiki, Nishioka Shingo, Sakurai Hirofumi, Imaichi Osamu, Iwayama Makoto, and Takano Akihiko. Extracting terms by a combination of term frequency and a measure of term representativeness. *International journal of theoretical and applied aissues in specialized communication*, Vol. 6, No. 2, pp. 211–232, 2000.

[9] Yang Yiming and O. Pedersen Jan. A comparative study on feature selection in text categorization. In *Proc. of ICML-97*, pp. 412–420, 1997.

[10] 若木裕美, 正田備也, 高須淳宏, 安達淳. 検索語の曖昧性を解消するキーワードの提示手法. *DBSJ Letters*, Vol. 4, No. 2, pp. 41–44, 2005.

[11] 若木裕美, 正田備也, 高須淳宏, 安達淳. 検索語の曖昧性を解消するキーワードの提示手法. 情報処理学会研究報告「データベースシステム」, 第 137 巻, pp. 269–276, 2005.

[12] 松尾豊, 石塚満. 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム. 人工知能学会論文誌, Vol. 17, pp. 213–227, 2002.

[13] 相澤彰子. 語と文書の共起に基づく特徴度の数量的表現について. 情報処理学会論誌, Vol. 41, pp. 3332–3343, 2000.

[14] 大澤幸生, ネルス E. ベンソン, 谷内田正彦. Keygraph: 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌, Vol. J82-D-I, pp. 391–400, 2 1999.