検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング

若木裕美[†] 正田備也[††]
高須淳宏[††] 安達淳[††]

現在の検索エンジンでは，入力される検索語に曖昧性があるため，結果のランキング出力の中に，質問者の期待していないノイズ文書が混在することが避けられない．そこで我々は，多様な内容を含む検索結果の中から，質問者の期待する内容に特化した検索結果を得られるような検索語を提示し，質問にあった曖昧性を解消するための手法を提案する．まず，特定のトピックに強く関係する単語を抽出するための単語の重み付け手法として，単語共起の統計に基づく定式化を提案する．特定のトピックに強く関係する単語を抽出することで，トピックを際立たせることが出来ると考える．そして，この手法によって抽出された単語を用いて単語クラスタリングを行って，検索結果の中に混在していた幾つかのトピックに分けて単語を質問者に提示することで，質問者が求めるトピックに相当する検索質問拡張用の単語を発見しやすくする．また，この提案手法の評価方法として，特定のトピックに強く関係する単語であることを示す定量的評価方法を提案し，被験者を用いた実験により，この評価方法が主観評価と強い相関があることを検証した．さらに，多種の既存の単語重み付け手法と比較して，提案手法の有効性を確認した．

Topic-oriented Term Extraction and Term Clustering for Query Focusing

Hiromi Wakaki,[†] Tomonari Masada,[††] Atsuhiro Takasu[††]
and Jun Adachi[††]

We often use search engines with queries consisting of few terms. Such short queries can be ambiguous, and it is difficult for users to select appropriate query terms. Accordingly, existing search engines may not avoid showing search results including documents irrelevant to users' query. Therefore, we will propose a method to suggest additional query terms which can retrieve appropriate search results focusing on users' intention and disambiguate the original query. First of all, we propose a formulation for term weighting calculation which can extract terms strongly related to a specific topic. The formulation is based on statistics of term co-occurrence. We believe that extracting terms strongly related to a specific topic can highlight various topics included in the search results. And then, making term clusters composed of the terms extracted by our method leads to isolate each of those topics. Showing those clustered terms can help users to discover appropriate terms for query expansion corresponding to users' intention. Moreover, we propose a pseudo evaluation metric for our term weighting method as an estimation of how strongly a term is related to a specific topic. Our results showed strong correlation between subjective evaluation by users and our pseudo evaluation metric. In our experiments with seven different sets of data, proposed term weighting method outperformed other several existing term weighting methods on both pseudo evaluation and subjective evaluation.

1. はじめに

現在の検索エンジンでは，結果のランキング出力の中に，質問者の期待していないノイズ文書が混在することが避けられない．そのため，ランキングされた結果の中には，様々な粒度の情報や多様な内容の文書が一次元的に並べられてしまうため，自分の要求に合致する内容を発見するのが困難である．また，質問者が検索語として適切な語を知っていれば良いが，それを見出すのが難しいことがしばしばある．そして，検索質問は 1〜2 語であることが多く[1)]，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．

そこで，検索語の曖昧性を解消するため，検索結果

† 東京大学大学院情報理工学系研究科電子情報学専攻
Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo
†† 国立情報学研究所
National Institute of Informatics

に含まれる多様なトピックを網羅しつつ，それらを明確に分離して表す単語群を提示する手法を提案する．我々は，一定の単語群が頻繁に同じ文書に出現することがトピックの現れだと考える．このため，ある単語が一定の単語群と頻繁に共起する場合，その単語は特定のトピックに強く関係していると考えられる．しかし，例えばストップ・ワードは，他の単語群と頻繁に共起するとはいえ，共起する単語の種類が多いため特定のトピックに関係するとは言えない．そこで,「特定の単語群とのみ頻繁に共起する」という単語の性質をTangibility と呼び，Tangibility の高い単語のみを抽出し，残りの単語は除去する．こうして除去される単語には，ストップ・ワードだけでなく，文書集合全体に関わる内容を代表するような単語も含まれる．従来の特徴語抽出法には，このような単語を抽出することを目標とするものもある．しかし，本研究では，複数のトピックに同時に関係しうる単語は抽出しないようにする．なぜなら，このような単語を除去しておけば，残された単語同士を共起関係にしたがって結びつけるだけで，文書集合に含まれる多様なトピックに対応する単語の集合を取り出すことができるからである．

そこで，単語の重み付け手法の提案に加えて，本手法によって抽出された単語による，単語クラスタリングについても実験を行った．こうして，検索結果の中に混在していた幾つかのトピックに分けて単語を質問者に提示することで，質問者は自分が求めるトピックに対応する検索質問拡張 (Query Expansion) 用の単語を発見しやすくなる．また，提示された単語群が，検索したい分野に詳しくない質問者にとって未知である場合，質問者の学習支援へもつながる．

さて，提案手法によって単語のトピックとの関連度が正しく定量化できているかを評価する方法は，我々の知る限りでは存在しない．そこで，我々は，この提案手法の評価方法として特定のトピックに強く関係する単語であることを示す定量的評価方法を提案し，被験者を用いた実験によりこの評価方法が主観評価と強い相関があることを検証している．さらに，多種の既存の単語重み付け手法と比較して，提案手法の有効性を確認している．

本稿では，3 節において，Tangibility の高い単語を抽出するための定式化を提案し，得られた単語を用いたクラスタリングアルゴリズムについて紹介する．4 節では，提案手法に対して比較する手法の説明を行う．5 節では，実験に用いたデータの説明を行う．更に，6 節において，トピックに強く関連する語であるかどうかの評価方法を提案し，提案する評価方法の妥当性を被験者を用いて実験し確認する．また 7 節において，実験結果について詳説する．7.1 節では，提案する評価方法を用いて，Tangibility の高い単語を取り出せているかどうかの評価を，7.3 節では，被験者の主観的評価による単語の評価を，7.4 節では，抽出した単語から生成した単語クラスタがトピックのまとまりを持つかどうかの評価を行う．最後に，8 節で本稿のまとめと今後の課題について述べる．

2. 関 連 研 究

本研究では，トピックを際立たせるような単語を抽出し，その単語をクラスタリングすることで，文書集合に含まれるトピックを提示する．また将来的には，検索質問拡張支援に利用することを目的とする．

文書分類の研究においては，次元の削減のために特徴語選択を行うことが多い．その発展として単語クラスタリングがある[2]．しかし，これらの研究では，文書の特徴量としての単語を効率的に抽出することを目的としており，本研究のように利用者に結果を直接提示することは想定されていない．

利用者に提示する情報の形態という視点からは，文書クラスタを用いる研究が多く行われているが，単語クラスタのほうが情報を把握しやすいと考え，本研究では単語クラスタを用いる．文書は，その全体を通じて特定のトピックを表現するが，単語には，特定のトピックを要約的に表現するものがある．それゆえ，特徴語選択の研究は多い．一般的な単語の重み付け手法[3)2)] に限らず，単語の代表性を測る方法[4]，文書中のキーワードを探すための単語クラスタリング[5]，単語の共起の関係を抽出する KeyGraph[6)7)]，検索結果として得られる文書群の内容的把握を助ける特徴語グラフをもつ DualNAVI[8] など多くの研究がある．いずれも提案手法とは目的が異なる．4) では，単語の代表性の定量化のための研究であり，文書集合に含まれるトピックを選別する目的ではない．5) では，単語のクラスタリングをしているが，1 つの文書を対象としており，本研究のような文書集合中のトピック抽出とは異なる．さらに，KeyGraph や DualNAVI では，単語間の関係を可視化しているが，本研究では単語をクラスタリングするのみにとどまり，単語間の関係を詳細に提示する手法ではない．

このように多くの特徴語選択の手法があるため，検索結果の文書クラスタリングした後に，特徴語を選択してトピックを提示するという方法も考えられる．しかし，単語を抽出する前に，文書という枠の中で文書のクラスタリングが行われるため，文書中に混在する

かもしれないトピックを見つけることが難しくなる．

一方，検索のための単語を探す方法として，検索質問拡張では様々な手法が研究されている．RSV (Robertson's Selection Value)[9] のように，検索結果から適当な単語を抽出して検索質問拡張を行う手法が一般的である．この場合，検索質問にある曖昧性の解消，すなわちトピックの分離は考えない．なお，検索質問にある曖昧性の解消の手法として query splitting という考え方がある[10][11]．これは元の検索質問に対し複数の見方がある場合に，複数のサブクエリに分割するという方法である．また，適合性フィードバックでは，検索者の選択する文書を元に検索質問拡張を行うことを前提とするが，多くの文書を閲覧するのは煩雑な作業である．検索結果の中に複数のトピックが混在しているときに，検索質問を改善するのに効果的と思われる単語やその単語を含む文例をピックアップして提示されれば，検索者にとって有益であると考える．実際に，検索結果を利用した単語の抽出や分類の応用として，12) や 13) がある．12) では，検索結果中の単語の出現確率や共起確率を使って単語間の親子関係を生成し，メニュー形式で階層構造を表現している．また，13) では，検索結果の概要の中に頻出する単語やフレーズを分類名として利用して実際に検索結果を表示するシステムを構築している．

検索者は，一旦検索された文書の中から検索者の意図に合致する内容を表す単語を探して利用することで，検索質問を改善することが多い．この場合，検索者が自分の知識を使って単語を選ぶが，実際には，検索結果として得られた文書集合によって，曖昧性を解消できる単語は異なると考えられる．そこで本研究では，検索結果の文書集合の特徴を使って検索結果の曖昧性解消に有効だと思われる単語を提示することで，検索者が通常行っているこのような作業について支援できると考えている．

3. トピックを強く表す単語の抽出とそのクラスタリング

3.1 単語の重み付け手法の提案

(1) 記号の定義

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．同様に，単語 t_i が出現する文書において単語 t_j が出現しない確率を $P(\neg t_j|t_i)$ と書く．また，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．単語 t_i が出現する文書集合を $S(t_i)$ と書くことにする．

(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 TNG と呼ぶ．

まず，単語 t_j の出現のしやすさが，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．

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

式 (1) は，Kullback–Leibler Divergence という情報量を求める式の一部であるが，その差異については後述する．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 TNG を得る．

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

式 (2) によって単語の重み付けを行い，単語を順位付けする．ただし，頻度が低い単語の場合の data sparseness の問題を回避するため，本来，

$$P(t_j|t_i) = \frac{|S(t_i) \cap S(t_j)|}{|S(t_i)|} \tag{3}$$

で推定するところを，以下のような式によって Dirichlet smoothing[14] を行った．

$$P(t_j|t_i) = \frac{|S(t_i) \cap S(t_j)| + \alpha|S(t_j)|}{|S(t_i)| + \alpha|S|} \tag{4}$$

「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[15][16]．ただし，15)，16) では，

部分的な文書集合以外に，補正項として，全体の文書の集合における単語出現頻度の情報を必要とした（検索タスクを例にあげると，検索可能な文書全体と検索結果としての部分的な文書集合の二つが必要であった）．そこで，本稿の定式化では，以前の定式化をより洗練させた．すなわち，部分的な文書集合のみから計算できる式に変更し，また,「特定の語とのみよく共起する」ということをより忠実に式に表現した．したがって，全体的な文書集合を必要としないため，対象とする文書データについての制約がなくなっただけでなく，全体的な文書集合の網羅性の影響を受けない．

(3) **Tangibility の式の意味**

情報量のひとつに Kullback–Leibler Divergence(KLD) 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 の出現に，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， KLD は次の式で表される．

$$KLD(t_j;t_i) = P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} + P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)} \quad (5)$$

TNG のための式 (1) は，式 (5) 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．増える場合には，この項は 0 より大となる．

TNG 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 TNG 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，特定の単語について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．

3.2 提案手法により抽出された単語による単語クラスタリング

(1) **単語間の類似度**

単語クラスタリングにおいては，単語間の類似度をどのように定義するかが重要である．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．

$$Sim(t_i,t_j) = \frac{|S(t_i)\cap S(t_j)|}{|S(t_i)\cup S(t_j)|} \quad (6)$$

ただし，$|S(t_i)\cap S(t_j)| < 5$ のときは $Sim(t_i,t_j) = 0$ とした．

(2) **クラスタリングのアルゴリズム**

本研究で提案している単語の重み付けの順位の影響が出るクラスタリング手法として，Baker らが提案する Distributional Clustering[17)18)] を用いる（**表 1**）．

表 1 Distributional Clustering アルゴリズム．
Table 1 Distributional Clustering Algorithm.

1. 相互情報量の値によって単語をソートする．
2. 上位の M 個の単語を，M 個のクラスタとすることで初期化する．
3. M 個のクラスタのうちの 1 つに全ての単語が入るまで、次の手順を繰り返す．
i) 最も近いクラスタをマージし，$M-1$ 個のクラスタとする．
ii) ソートリストの中で次の順位の単語によって，新しいクラスタを作る．

ただし，Baker らは各単語に対する相互情報量の値によって単語をランキングしているが，本稿では，相互情報量の値の部分を，TNG などの単語重み付け手法と置き換える．クラスタ間の類似度 $Sim(C_1,C_2)$ は，

$$Sim(C_1,C_2) = \frac{s(C_1,C_2)}{s(C_1,C_1)\times s(C_2,C_2)} \quad (7)$$

とした．ただし，C_1，C_2 は単語クラスタであり，$s(C_1,C_2)$ は，

$$s(C_1,C_2) = \sum_{t_i\in C_1}\sum_{t_j\in C_2} Sim(t_i,t_j) \quad (8)$$

と定義する．そして，最も類似度の高いクラスタペアをマージした．

本研究では，$M = 10$，単語は上位の 100 語のみとした．また，すべてのクラスタ間の距離が無限大となる場合 (別々のクラスタに属する単語間で共起が全くない場合) は，マージしないことにした．

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[19)]．しかし，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，上記のクラスタリング手法のみ試した．

4. 比較手法

本研究で提案する TNG と比較するための単語の重み付け手法として，相互情報量 (Mutual Information: MI)[3]，KLD，χ^2 [2]，Robertson's Selection Value(RSV)[9][20] の 4 つを対象とした．このうち，MI，KLD，χ^2 は，TNG と同様に単語の共起の情報を使い，ある単語 t_i が出現することで，他の単語 t_j の出現しやすさがどの程度変化するかを表す．このため，単語 t_i についての重みは，次のような式とする．

$$W(t_i) = \sum_j X(t_j; t_i) \tag{9}$$

ただし，$X(t_j;t_i)$ は，それぞれ $MI(t_j;t_i)$，$KLD(t_j;t_i)$，$\chi^2(t_j;t_i)$ に置き換えるものとする．また，この 3 手法については，TNG と同様にスムージングを行った．

4.1 MI

$MI(t_j;t_i)$ は，次のような式で書ける．

$$\begin{aligned} MI(t_j;t_i) = & P(t_i)\{P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} \\ & + P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)}\} \\ & + P(\neg t_i)\{P(t_j|\neg t_i)\log\frac{P(t_j|\neg t_i)}{P(t_j)} \\ & + P(\neg t_j|\neg t_i)\log\frac{P(\neg t_j|\neg t_i)}{P(\neg t_j)}\} \end{aligned} \tag{10}$$

4.2 Kullback–Leibler Divergence

$KLD(t_j;t_i)$ は，式 (5) を用いる．

4.3 χ^2

$\chi^2(t_j;t_i)$ は，次の式で表される．

$$\begin{aligned} \chi^2(t_j;t_i) = & \frac{\{P(t_j|t_i)-P(t_j)\}^2}{P(t_j)} \\ & + \frac{\{P(\neg t_j|t_i)-P(\neg t_j)\}^2}{P(\neg t_j)} \\ & + \frac{\{P(t_j|\neg t_i)-P(t_j)\}^2}{P(t_j)} \\ & + \frac{\{P(\neg t_j|\neg t_i)-P(\neg t_j)\}^2}{P(\neg t_j)} \end{aligned} \tag{11}$$

4.4 RSV

RSV 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．$RSV(t)$ は以下の式で定義される．

$$\begin{aligned} RSV(t) = & (\frac{rdf}{R} - \frac{df}{N}) \\ & \times \left\{ k \times \log\frac{N}{df} + (1-k) \times \log\frac{\frac{rdf+0.5}{R-rdf+0.5}}{\frac{df-rdf+0.5}{N-df-R+rdf+0.5}} \right\} \end{aligned} \tag{12}$$

ただし，rdf を部分文書集合 S 中で単語 t を含む文書数，R を文書集合 S に含まれる文書数，df を全文書集合 U 中で単語 t を含む文書数，N を文書集合 U に含まれる総文書数，k をパラメータとする．

今回の実験では $k = 0.5$ とした．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．例えば検索対象となる文書全体のような全体集合) が与えられていなければならない．

5. 実験用データ

5.1 データの種類

本研究の目的に合うデータとしては，複数のトピックを含む文書集合において，混在するトピックへの分類が存在しているデータを使いたい．分類用のデータであれば元々分類が付いているが，検索用のデータであっても各検索課題に対して適合文書集合が存在するため，分類データの場合と同様に幾つかのトピックが存在するデータとみなすことができる．そこで，分類用のデータと検索用のデータを両方利用した．また，データは日本語・英語の両方で実験を行った．

分類用のデータの場合，あるいは，ディレクトリ構造になった Web 上の文書集合を利用する場合，分類タスクにおけるカテゴリを本論文でのトピックとみなす．本研究では，各データ中に含まれる分類のうちで，最も文書数の多い分類のものを 3 つ混ぜて，擬似的にトピックが混ざった文書集合を生成しておく．そして，ある単語重み付け手法が文書の分類情報を全く使わずに，文書集合に含まれる単語の中から特定の分類に強く関係する単語を抽出できるとき，その手法は良い手法だと言える．

ただし，比較手法のうち RSV だけは，検索質問拡張用の単語重み付け手法であるため，3 つの分類が混合された部分的な文書集合だけでなく，それ以外のデータ全てを含む文書集合を必要とする式である．そこで，単語抽出の対象となる 3 つの分類だけでなく，全部の文書を RSV の全体文書集合として使用した．

更に，検索用のデータの場合，それぞれの検索課題についての適合文書集合が存在する．1 つの問い合わせに対する適合文書集合をトピックとみなし，分類用のデータと同様に扱う．また，検索対象として与えられる文書全体が存在するので，RSV の場合にも問題なく利用できる．

5.2 使用したデータ

実験で用いたデータは、**表 2**，**表 3** にある 7 種類である．各データを分類のついた文書集合とみなし，最

も多い分類 3 つを混ぜて実験用データとした．各データの言語，元々のタスクの形式，使用した 3 分類に含まれる文書数は表 3 の通りである．また，各データについて使用した 3 つの分類名の一覧は，表 2 であり，表 2 中の A, B, C は混合する 3 つの分類を表す．

(1) **NTCIR3，NTCIR4**[21)22)]

NTCIR3，4 の web タスクのために用意された，検索質問ごとの適合文書集合を利用した．

(2) **Dmoz**☆

Web 上にあるディレクトリである Dmoz のデータの中から，Science の下のディレクトリを利用した．

(3) **Reuters**

文書分類タスクに使われるデータの 1 つである Reuters-21578 を使用した．

(4) **産経スポーツ**☆☆

Web 上にある産経スポーツのバックナンバーを利用した．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[21)] 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．☆☆☆

(5) **Newsgroup20**

文書分類タスクに使われるデータの 1 つである．いずれの分類も 1000 文書程度で構成されている．

(6) **NTCIR-CLIR**[23)]

検索用の課題で，NTCIR3 において実施されたものを使った．言語横断検索用で，毎日新聞の英語版のデータである．

5.3 データの特性と妥当性

本研究の実験では，人工的にトピックが混在するデータを合成して使用した．そのため，実際の検索過程で得られるものとは異なるデータである．実際の検索結果では含まれるトピックの粒度が異なることが予想され，それらのトピックを適切に決めることが難しいため，客観的かつ定量的に評価する目的に合わない．また，検索結果の中に含まれる曖昧性解消の評価に使うことのできる正解データは，我々の知る限りでは存在しない．このため，トピックの分離が正確になされていると思われるデータを利用して，人工的にトピックが混在するデータを合成して使用した．

本実験では，検索語に関連した幾つかのトピックの分離を想定している．実験で用いるデータは，混合する幾つかのトピックの区別ははっきりしているが，同一のトピックに関連した内容という意味では，現実の問題に即するような曖昧性を含むデータとなっている．まず，産経スポーツのニュース記事では，すべてスポーツの内容であるだけでなく，日本の野球とメジャーリーグは同じ野球の内容を扱っており，同一トピックに関連する．NTCIR3 のデータでは，憲法第九条に関する文書と著作権に関する文書が混じっており，すべて法律に関連する．さらに Dmoz では，Science の下のディレクトリを用いているので，同一トピック内の内容であるといえる．同様に Newsgroup20 でも，すべて politics に関する文書である．

実験で使用したデータは，産経スポーツのデータ以外は，いずれも多くの研究において利用されているデータであり，トピックの分離が客観的に信頼されていると考えられる．さらに，産経スポーツのデータも人手できちんと分類されたものである．また，本研究では，実システムの構築ではなく，手法の妥当性を客観的に評価することを目的としている．このため，人工的なデータで現実の状況を近似して実験を行った．

5.4 データの整形

日本語のデータは，MeCab[24)] を用いて形態素解析を行った．辞書は ipadic-2.5.1[25)] である．また，英語のデータは，Porter のステミングを行い，一般的なストップワード 約 100 語を除去した．いずれのデータでも，DF の高い上位の 1000 語について実験を行った．

6. 単語の評価方法の提案（単語のラベル付け推定と単語のトピック関連度の推定方法について）

6.1 単語のラベル付け推定とトピック関連度の推定が必要な理由

本節では，文書分類のラベルを用いて，各単語に対して人間が付与するであろう「分類ラベル」と，「トピックとの関連度」を測る方法を提案する．また，被験者を用いた主観的評価により，提案する単語のラベル付け推定の方法が人間の付けるラベル付けとほぼ一致すること，そして，各単語におけるトピックとの関連度がその単語に対して一致して同じラベルを付けた人数と相関があることを実験で確認していく．

一般に，文書クラスタリングにおいては，正解データの用意されているデータ・セットが存在するため，クラスタを生成した後に精度を測定することが可能であ

☆ http://dmoz.org/

☆☆ http://www.sanspo.com/

☆☆☆ RSV の式 (12) 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．

表 2 使用したデータと，その中の分類名または検索質問.
Table 2 Data we used, and class names or query terms within the data.

データ	A	B	C
NTCIR3	「憲法，第九条，解釈」	「京都，寺，神社」	「著作権，デジタルコンテンツ，ネットワーク」
NTCIR4	「競馬，血統」	「哲学，存在論」	「中国経済，社会主義，市場」
Dmoz	Math	Chemistry	Astronomy
Reuters	earn	acq	crude
産経スポーツ	日本の野球	メジャーリーグ	サッカー
Newsgroup20	talk.politics.guns	talk.politics.mideast	talk.politics.misc
NTCIR-CLIR	「Give information regarding protests against nuclear power.」	「Articles relating to President Kim Dae-Jung's policy toward Asia」	「Incidents relating to religious thought about doomsday, or the end of the world.」

表 3 実験に使用した各データの言語，形式，文書数.
Table 3 Languages, purposes, and numbers of the documents we used for our experiment.

データ	言語	タスクの形式	文書数 (表 2 中の A+B+C)	全文書数
NTCIR3 web	日本語	検索	1108(476 + 282 + 350)	10253810
NTCIR4 web	日本語	検索	2113(643 + 722 + 748)	10253810
Dmoz	英語	分類	21089(8935 + 5584 + 6570)	63300
Reuters	英語	分類	6615(3845 + 2362 + 408)	9494
産経スポーツ	日本語	分類	3519(1233 + 757 + 1529)	10257329
Newsgroup20	英語	分類	3000(1000 + 1000 + 1000)	19955
NTCIR-CLIR	英語	検索	209(135 + 50 + 24)	12723

る．しかし，本研究では単語のクラスタリングを目的としているため，文書にラベルが付いたデータではなく，単語にラベルが付いている正解データが必要となる．だが，単語にラベルが付与されたクラスタリング用の正解データは存在しない．加えて，多言語データを使った実験の評価には，各言語を母国語とする被験者を用意する必要があるが，これは困難である．したがって、人間が付与するであろうラベルを推定する方法が不可欠である．更に，本研究では，特定のトピックに強く関連する単語の重み付け手法を新しく提案している．そこで，この単語の重みの大小が，その単語におけるトピックとの関連度の強弱と対応しているかどうかの評価も，同様に人間が判断する必要がある．人間がラベルを付けることは可能だが，あらゆる実験用データについて被験者にラベル付けを行わせ，各々の単語に対して一致して同じラベルを付けた人数を調査することは困難である．このため,「分類ラベルの推定方法」「トピックとの関連度の推定方法」の二つを提案し，人間による主観評価実験との照合を行って，その妥当性を示す.

6.2 評価の推定方法

分類の付いている文書データを利用し，単語が関連するであろうトピックに相当するものとして，文書に付与された分類を想定することにする．本稿における実験では，3つの分類のいずれかが各文書に付与されたデータを用いる．ここで，トピックとの関連度 (topic partiality) の推定値 $TP(t_i)$ を，次の式で求める．

$$TP(t_i) = \frac{DF(t_i)}{N} \times K(t_i) \tag{13}$$

ただし，

$$K(t_i) = \sum_j p_j(t_i) \log \frac{p_j(t_i)}{q_j} \tag{14}$$

とする．$K(t_i)$ は，Kullback–Leibler Divergence である．確率 $p_j(t_i)$ $(j = 1, 2, 3)$ は，単語 t_i を含む文書のうち分類 $j(j = 1, 2, 3)$ に属するものの割合である．また，確率 $q_j(j = 1, 2, 3)$ は，全文書数に対する，それぞれの分類 $j(j = 1, 2, 3)$ に所属する文書の割合である．N は，全文書数である．

つまり，文書に付与されている分類 $(j = 1, 2, 3)$ の各割合 (q_1, q_2, q_3) に対して，各単語が出現する頻度 $(p_1(t_i), p_2(t_i), p_3(t_i))$ が偏っているほど，式 (14) の値は大きくなると考えられる．ただし，出現頻度が低い単語においては，いずれかの分類へ偏りやすいため，$\frac{DF(t_i)}{N}$ を $K(t_i)$ に掛けることで補正している．

各単語がいずれのトピックであるかは，

$$TPL(t_i) = \arg\max_j p_j(t_i) \log \frac{p_j(t_i)}{q_j} \tag{15}$$

で推定する．つまり，$p_j(t_i) \log \frac{p_j(t_i)}{q_j}$ を最大にする $j = k$ が，単語 t_i に付与されるべきラベルであるとする．また，トピックへの関連度は，分類 k に対して $TP(t_i)$ であると推定する．この妥当性は 6.4 節で

示す．

6.3 主観評価実験

主観的評価のための被験者は，20 代〜30 代の日本人の男女 20 名である．日本人を対象としたため，実際にチェックしてもらうデータは日本語のデータを使用した．表 2 の中の産経スポーツ，NTCIR3，NTCIR4 の 3 種類のデータと分類を使用し，各データにおいて 3 種類の分類ラベルを混合した文書データの中で単語の重み付けを行い，上位にあがった各単語だけについて元々文書分類として用意されたラベルを付けてもらう．単語は，提案手法である TNG，比較手法である MI と RSV，以上 3 種類の重み付けによってランキングされたそれぞれの上位 100 語（重複する単語があるため合計 200 語強）である．3 つの各データに対してそれぞれ単語データセットが出来るが，その中の単語の順番をランダムに並べ替えたものを被験者に提示する．

被験者には，産経スポーツ，NTCIR3，NTCIR4 から得られた各単語データセットに対して，各データに含まれる 3 つの分類ラベル（表 2 参照）を与え，1 つ 1 つの単語に 3 つのうちのいずれかのラベルを付けてもらった．ただし，いずれの分類であるかを判断できない（複数のものが対応するか，単語自体が曖昧でどの分類か判断できない）場合と，単語自体を知らない（あるいは，意味が分からない）場合の 2 つのラベルも別に用意し，個人の判断基準に任せてラベルを付けてもらった．

全ての単語について，20 人のうち何人が一致して同じラベルを付けたかによって，各単語のトピックへの関連度の強さを調査する．このため，データは単語数分プロットされる．

6.4 単語の評価の推定方法と主観的評価の関係

式 (13) によって推定されたラベル付けとトピックへの関連度を，被験者を用いた主観的評価によって検証した．特に NTCIR3 のデータに関しては末尾に配置した**図 6** に，それ以外のデータも含めた相関係数は**表 4** に示した．ただし，相関係数は，ピアソンの積率相関係数を用いた．また，$TPL(t_i)$ によるラベル付けと被験者によるラベル付けの単語数の一覧を**表 5** に示した．

まず，図 6 は，式 (13) によってラベル付けされた単語の $TPL(t_i)$ のラベルごとの図である．$TPL(t_i)$ のラベルが，a)「憲法，第九条，解釈」b)「京都，寺，神社」c)「著作権，デジタルコンテンツ，ネットワーク」であるとき，各単語 t_i に対して，A，B，C，判断できない，知らない，の 5 つのそれぞれのラベルを何人の被験者がつけたかを表す．縦軸が同じラベルを付けた被験者数（最大となるラベルではなく，すべてのラベルについて表示），各点が単語，各系列はユーザが付与したラベルを示す．図中において，系列の記号が黒く塗られているものは，$TPL(t_i)$ のラベルと被験者の主観的評価のラベルとが同じであった単語である．反対に，系列の記号が白抜きになっているものは，$TPL(t_i)$ のラベルと異なるラベルが被験者の主観的評価によって付与された単語である．また，系列記号×は，いずれのラベルか判断できないとされた曖昧性の高い単語，系列記号＋は，被験者がそもそも知らない単語であったことを表す．さらに，図 6 において，系列の記号が黒く塗りつぶされている単語が，もし正の相関があれば，$TP(t_i)$ の値によって人間の感じるトピックとの関連度が正しく値が評価できているといえる．また逆に，×の系列が負の相関を持てば，$TP(t_i)$ の値が低いほど曖昧性が高いといえる．

図 6 において，多数の被験者が $TPL(t_i)$ のラベルと同じラベルを付与していることが分かる．このことから，$TPL(t_i)$ のラベルは人間の付けるラベルを推定できることが言える．また，いずれの実験データにおいても，$TP(t_i)$ の値と，同じラベルを付けた被験者数に強い正の相関が見られる (図 6)．また，判断できないとされた単語については，$TP(t_i)$ の値と負の相関が見られ，値が低いほど曖昧性が高くなっていることもいえる．

同様に，表 4 では，太字の相関係数は正しくラベルが付けられていた場合で，この相関係数が高いと $TP(t_i)$ と $TPL(t_i)$ が正しく推定できているといえる．また，「判断できない」というラベルにおいて負の相関があるとき，単語の曖昧性も測れており，$TP(t_i)$ の推定が良いといえる．さらに，相関係数の統計的有意性を無相関検定によって確認した．相関係数の値の右側に* 印がついている場合は，5 %の有意水準で有意となった相関係数である．実際，表 4 では，$TP(t_i)$ の値と，同じラベルを付けた被験者数に強い正の相関が見られ，逆に，判断できないというラベル付けがされた単語については，$TP(t_i)$ の値と負の相関が見られる．

表 5 の太字部分は，$TPL(t_i)$ が正しくラベルを付与した単語数，それ以外は間違えたラベルを付与した単語数である．表 5 中 (1) を見ると，被験者 20 人中 10 人（過半数）以上が付与したラベルと異なるラベルを $TPL(t_i)$ が付与することはなかった．すなわち，大多数の人間が見てどのトピックであるかが分かる場合，$TPL(t_i)$ のラベル付けは間違わない．しかし，いずれのラベルか判断できないとした人が多かった単語

表 4 被験者によるラベル付けと推定値 $TP(t_i)$ との相関係数．各ラベル A，B，C は，表 2 を参照．
Table 4 Correlation coefficient between subjective evaluation by users and our pseudo evaluation metric $TP(t_i)$.

a) 産経スポーツのニュース記事

(下) 被験者の付けたラベル (右)$TP(t_i)$ の付けたラベル	A	B	C
A	**0.85***	0.0039	-0.15
B	-0.034	**0.52***	0.26*
C	0.058	-0.28	**0.76***
判断できない	-0.79*	-0.44*	-0.75*
知らない	-0.24	-0.052	-0.019
単語数	110	23	92

b)NTCIR3

(下) 被験者の付けたラベル (右)$TP(t_i)$ の付けたラベル	A	B	C
A	**0.65***	-0.25	0.12
B	-0.16*	**0.69***	-0.051
C	-0.15*	-0.27	**0.14**
判断できない	-0.58*	-0.63*	-0.21
知らない	-0.317	-0.40	0.12
単語数	163	13	30

c)NTCIR4

(下) 被験者の付けたラベル (右)$TP(t_i)$ の付けたラベル	A	B	C
A	**0.77***	-0.18	-0.17
B	-0.23	**0.77** *	-0.32
C	-0.16	0.11	**0.84***
判断できない	-0.75*	-0.73*	-0.81*
知らない	-0.25	-0.33	-0.34
単語数	63	105	37

もあった．例えば同じ単語に対して，A に 4 人，B に 2 人，C に 0 人，判断できないに 1 4 人がラベルを付与したとすると，A を付与した人が他のラベルに比べて多いので $TPL(t_i)$ が A を付与すれば良い．しかし，あいまいでいずれのラベルか判断できないと多くの人が考えているので，間違いやすく判定が難しいことが想定される．このような場合であっても，すなわち，表 5 中 (2) で他のラベルよりもあるラベルにつけた人数が多い場合についての比較でも，そのラベルと同じラベルを $TPL(t_i)$ が付けた場合は，異なるラベルを付ける場合に比べて圧倒的に多い．このことから $TPL(t_i)$ のラベル付けも正しく行われていると考えられる．以上より，本稿では，$TP(t_i)$ を主観的評価の推定値，$TPL(t_i)$ を推定ラベルに用いることとする．

7. 実験と結果

7.1 $TP(t_i)$ を用いた各データにおける単語の評価

表 2 の各データについて，TNG，MI，KLD，χ^2，RSV の各手法の性能を比較した．ただし，TNG，MI，KLD，χ^2 のスムージング用のパラメータ α は，それぞれ 5 つのパラメータ ($\alpha = 0.05, 0.1, 0.3, 0.5, 1.0$) について実験している．Reuters のデータについて得られた結果は，**図 1** である．

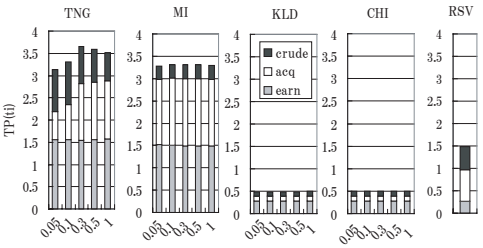


図 1 各手法によって，Reuters のデータから抽出した単語の $TP(t_i)$ の値による評価．
Fig. 1 Evaluation of term weighting methods by $TP(t_i)$. The terms are obtained from Reuters documents.

図 1 を見ると，$\alpha = 0.3$ の場合が高くなっているが，他のデータにおいても，概して $\alpha = 0.3$ の場合に $TP(t_i)$ の値が高くなったため，$\alpha = 0.3$ についてのみ図示する（**図 2**）．すべてのデータの実験結果に対して，各手法による上位 100 語の $TP(t_i)$ 値について，$TPL(t_i)$ によって推定された分類ごとの割合が分かるようにまとめたものが図 2 である．ただし，同一データを使用した際の手法間の比較であり，異なるデータ間の比較ではない．

いずれのデータにおいても，TNG が比較的高い $TP(t_i)$ を得ている．また，図 2 で A，B，C がいずれもバランス良く現れていることより，TNG は各トピックに対応する単語の抽出も良いことが分かる．ここで重要なのは，網羅的に各トピックに関連する単語が得られている点である．さらに，表 3 の文書数から，各トピックの文書量に偏りがあることがわかる．特に Reuters では，A と C で 9 倍程度の差がある．しかし，図 2 の TP 値には，文書量に依存する影響は出ておらず，文書量に偏りがあったとしても，網羅的に各トピックに対応する単語が得られているため問題はない．

7.2 3つ以上のトピックを混合したデータにおける $TP(t_i)$ による評価

本実験においては 3 つのトピックを混合した人工的なデータを使うことを基本としているが，その妥当性について検討する．複数のトピックが混ざったデータに関して，それらの分離を考える．このとき，二つのトピック（例えば A と B）を混ぜた場合だと，A と B を分けるという手法と，A と A でないものを分け

表 5 被験者の付けたラベルと $TPL(t_i)$ のラベル付けによる単語数の一覧
Table 5 Number of terms labeled by users and by TPL.

(下) 被験者の付けたラベル		$TPL(t_i)$ でつけたラベル								
		産経スポーツ			NTCIR3			NTCIR4		
		A	B	C	A	B	C	A	B	C
比較 (1) 10 人以上が 次のラベルを付与	A を付与	**21**	0	0	**45**	0	0	**33**	0	0
	B を付与	0	**5**	0	0	**9**	0	0	**20**	0
	C を付与	0	0	**29**	0	0	**23**	0	0	**21**
比較 (2) 次のラベルを付与した人が 他のラベルより多い	A を付与	**56**	9	14	**109**	0	2	**53**	7	3
	B を付与	2	**9**	1	2	**13**	0	2	**63**	5
	C を付与	6	4	**47**	5	1	**28**	2	13	**32**
TPL の各分類に含まれる語数		110	23	92	163	13	30	63	105	37

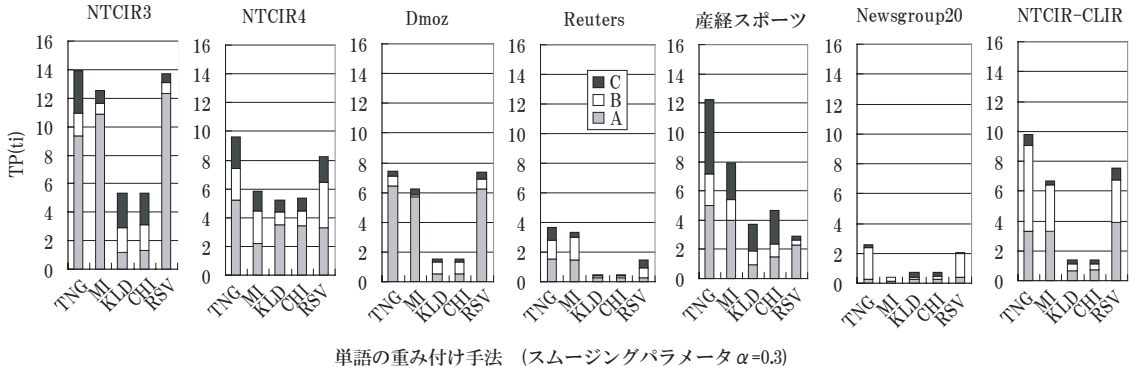


図 2 各単語重み付け手法の $TP(t_i)$ による評価．7 つのデータを使用．系列 A，B，C は，表 2 中の分類を参照．
Fig. 2 Evaluation of term weighting methods by $TP(t_i)$. We used seven different data.

るという手法の違いが分からない．そこで，最小限の設定として 3 つのトピックを混合することを採用した．また，被験者によるラベル付けも行うため最小限の設定とした．

しかし，実際の検索結果では，3 つ以上のトピックが存在することも考えられる．そこで，3 つ以上の場合についても NTCIR3 のデータを用いて実験を行った．表 2 のラベル A，B，C に加えて，D を「ブルーベリー，アントシアニン，視力」(188 文書)，E を「三国志，ゲーム，題材」(167 文書) とする．このとき，A-D の 4 つを混ぜた場合と A-E の 5 つを混ぜた場合について実験を行った結果が，**図 3**a)b) である．

3 つ以上の分離の実験 (図 3) でも，3 つのときの結果 (図 2 中の NTCIR3) と似た傾向が見られた．特に TNG については，4 つのトピックのときも，5 つのトピックのときも，最も $TP(t_i)$ 値が高く，かつ，すべてのトピックに対応する単語が得られており良い結果であった．この結果から，3 つのトピックに限定して実験を行うことで，基本的なパフォーマンスについて測ることができていると考えられる．

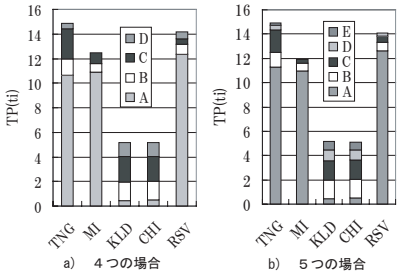


図 3 NTCIR3 のデータにおいて 3 つ以上のトピックを混合した場合の $TP(t_i)$ による評価．ただし，a) 4 つのトピックを混合した場合，b) 5 つのトピックを混合した場合．
Fig. 3 Evaluation of term weighting methods by $TP(t_i)$ using pseudo data containing more than 3 topics of NTCIR3. a)pseudo data containing four topics and b) pseudo data containing five topics.

7.3 被験者の付与したラベル付けによる評価

TNG，MI，RSV の 3 つの手法によって上位にランク付けされた単語について比較を行った．TNG と MI

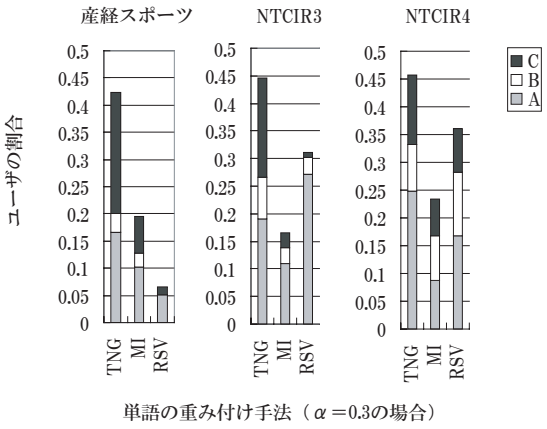


図 4 被験者が実際に付与したラベル付けによる単語の評価.
Fig. 4 Evaluation of terms by subjective evaluation.

については，スムージング用のパラメータを $\alpha = 0.3$ としてある．

被験者が付与したラベル付けによって，産経スポーツの記事，NTCIR3，NTCIR4 のデータを用いて抽出した単語に関する評価が，**図 4** である．各単語について，全被験者に対する，同じラベルを付けた被験者数の割合を，各単語のトピックとの関連度として評価した．すなわち，各単語のトピックへの関連度 $UT(t_i)$ は，次のように表せる．ただし，単語 t_i に対してラベル L_j を付与した被験者数を $U(L_j, t_i)$ とする．

$$UT(t_i) = \max_j \frac{U(L_j, t_i)}{20} \tag{16}$$

また，UT を最大にする分類 L_j を，単語 t_i の分類とみなした．そして，TNG，MI，RSV の各手法で得られた上位の単語 100 語に対して，分類ごとに，その分類へラベル付けされた単語のトピックへの関連度の総計を求め，グラフに表したのが，図 4 である．

$TP(t_i)$ による評価（図 2）に比べると，TNG への評価が高いことが分かる．また，図 4 において，TNG では被験者からの評価によるトピックへの関連が強い単語が多いだけでなく，各分類に網羅的に対応する単語が抽出できていることが分かる．各分類から網羅的に単語が抽出されないと，単語をクラスタリングした際にどのクラスタにも対応しないトピックが出てくることが予想されるため，網羅性が高いことも重要な特徴の 1 つである．また，通常の検索では，検索結果の上位が 1 つのトピックで占められてしまう問題がある．しかし，クラスタが幾つかのトピックを網羅していれば，この問題も回避できる．

7.4 単語クラスタの評価

提案手法により抽出された単語による単語クラスタにおいて，正しくクラスタリングがなされているかどうかの検証を行う．クラスタリングのよさを測る方法として，Microaveraged Precision[26)2)] というものがある．Microaveraged Precision では，単純に各クラスタの精度の平均を取るのではなく，クラスタのサイズが大きいものの影響が出やすいようになっている．

まず，各クラスタの精度を $Prec(C_j)$，$class_k$ をクラスタ C_j の中で一番多かったラベルとするとき，$Prec(C_j)$ を次のように定義する．

$$Prec(C_j) = \frac{\sum_{t_i \in class_k} TP(t_i)}{|C_j|} \tag{17}$$

このとき，Microaveraged Precision の計算方法に準じて，$TP(t_i)$ 値の大小を反映した下記のクラスタリング評価式 MP を提案する．

$$MP = \frac{\sum_j \sum_{t_i, j \in class_k, j} TP(t_i)}{\sum_j |C_j|} \tag{18}$$

MP を計算した結果が，**図 5** である．TNG，MI，KLD，CHI については，$\alpha = 0.3$ のときの結果を図示した．クラスタリング後の結果においても，TNG が高い MP 値を保持している．これは，単語クラスタリングをした結果において，各クラスタの精度が高く，かつ，含まれる単語がそれぞれのトピックに強く関連した単語であることを示しているといえる．

実際の単語クラスタの例は，**表 6** である．表 6 は，a)TNG，b)MI，c)RSV のそれぞれの手法により，NTCIR3 の 3 つの検索質問「憲法，第九条，解釈」,「京都，寺，神社」,「著作権，デジタルコンテンツ，ネットワーク」の適合文書集合を混合したデータを用いて，抽出した上位 100 語による単語クラスタリングの結果である．$M = 10$ としているため，10 個のクラスタが生成された．TNG の方が MI や RSV よりも具体性の高い単語で形成されており，各クラスタのトピックがはっきりしていることが確認できる．また，図 2 を見ると，RSV は MI よりも高い TP 値を得ているが，A，B，C のバランスが他のいずれの手法に比べても悪い．この結果，表 6 の単語クラスタでは，ほとんど「憲法，第九条，解釈」のクラスタしか出来ない結果となってしまっている．このように，具体性の高い単語で構成されていることだけでなく，トピックが網羅できているという点でも、TNG は他手法に比べて良い結果であった．

各単語クラスタと対応するトピックを調べると，表 6a)TNG の場合には，1 番目のクラスタが「京都，寺，神社」，2-3 番目のクラスタが「著作権，デジタルコンテンツ，ネットワーク」，4-9 番目のクラスタ

が「憲法，第九条，解釈」に関連したクラスタであるように見える．このようにトピックに対応するクラスタの数に偏りが見られるが，すべてのトピックを網羅している．今後，クラスタリングの手法の改善などを通じて，クラスタ数の偏りは改善していけるものと考える．

8. おわりに

本研究では，新しい単語の重み付け手法として TNG を提案した．1 つの文書の中で共起しやすい単語群をトピックとみなし，特定の単語とのみ強く共起するということの定式化を提案した．これにより，特定のトピックに強く関わる単語を抽出することが可能となる．更に，各単語の特定のトピックへの関連度を推定する式を新たに定義し，また，各単語のラベル付けを推定する方法も提案した．各単語のトピックへの関連度推定および各単語のラベル付け推定の有効性については，被験者を用いた主観的評価実験によって検証を行った．そして，提案手法の評価としては，各単語のトピックへの関連度推定を用いた単語の評価と，被験者による主観的評価の双方から実験を検証した．このことにより，提案手法では，特定のトピックへ強く偏る単語を抽出することができるということを確認した．また，各単語のトピックへの関連度推定を用いたクラスタリングの評価式により，提案手法によって抽出された単語が特定のトピックを表す単語クラスタを形成することを確認した．今後は，実用的なシステムを目指して，各単語クラスタに含まれる単語を用いた検索質問拡張への応用に取り組む予定である．

表 6 a)TNG， b)MI，c)RSV によって抽出した単語を用いた単語クラスタ．
Table 6 Examples of term clusters made by a)TNG, b)MI and c)RSV.

a)TNG

単語群
神社 京都 寺 堂 平安 境内 祭 宮 京 行事
コンテンツ デジタル 著作 音楽 配信 コピー ネットワーク データ 不正 画像 検索 サイト システム
電子 流通 ソフトウェア インターネット 普及 保護 技術 販売 ビジネス アクセス ソフト 町 サービス
武力 報復 テロ 自衛隊 小泉 国連 犠牲 根絶 戦闘
軍事 自衛 軍 憲法 戦争 市 内閣 安保 国民 米 政権
後方 平和 行使 集団 党 首相 攻撃 行動 決議 政党
開発 派遣 選挙 自民党 保障 やる 民主 国会 政治 政府
世論 日本国 議員 九 事態 発言 与党 ブッシュ 反対 戦後 改革 総理 しれる
院 同盟 賛成 湾岸 話 主義 危機 軍隊
輸送 感じる いける

b)MI

単語群
神社 京都 寺 後 市 せる
政府 国民 国会 力 出る 立場 行使 求める 言う 持つ
憲法 戦争 私 平和 状況 協力 軍 ところ とる たち
問題 それ これ 主義 認める 受ける 行為 政策 強い 出す 今回 明らか 参加
国 何 点 くる 以上 いく 人 自由 自衛 判断 主張 安全 数
経済 政治 ない アメリカ 思う 基本 場合 軍事 反対 対応 民主 意味 集団
社会 国際 行動 米 自衛隊 得る 解決 活動 しまう 首相 改正 とき 国連
考える 性 関係 必要 責任 議論 国家 結果 大きい 事態
コンテンツ 著作 デジタル 支援 化 間 可能 委員 具体 いう
本 制度

c)RSV

単語群
憲法 戦争 軍事 平和 テロ 日本 問題 立法 アメリカ 九
権 権利 世界 事態 改憲 主張 時代 保障 守る られる 保護 院 社会事実 的 これ
自衛 国連 武力 行使 小泉 紛争 許す 決議 協力 いう
自衛隊 軍 米 攻撃 集団 国際 軍隊 解決 報復 ため 支援 措置 それ
国会 政府 党 国 安保 反対 政権 内閣 違反 同盟
条 国民 政治 法 外交 放棄 認める 行動 議員 三
首相 行為 自民党 改正 法案 民主 防衛 与党 事件 十 立場
神社 京都 寺
主義 議論 日本国 解釈 条約 国家 侵略 米国 上 著作 重要 自由 政策
手段 法律 政党 諸国

参 考 文 献

1) Jansen, B. J., Spink, A., Bateman, J. and Saracevic, T.: Searchers, The Subjects They Search, And Sufficiency: A Study Of A Large Sample Of Excite Searches, *1998 World Conference on the WWW and Internet* (1998).
2) Sebastiani, F.: Machine learning in automated text categorization, *ACM Computing Surveys*, Vol.34, No.1, pp.1–47 (2002).
3) Yiming, Y. and Jan, O. P.: A comparative study on feature selection in text categorization, *Proc. of ICML-97*, pp.412–420 (1997).
4) Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. and Takano, A.: 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).
5) 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).
6) Ohsawa, Y., Nels, E., and Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence

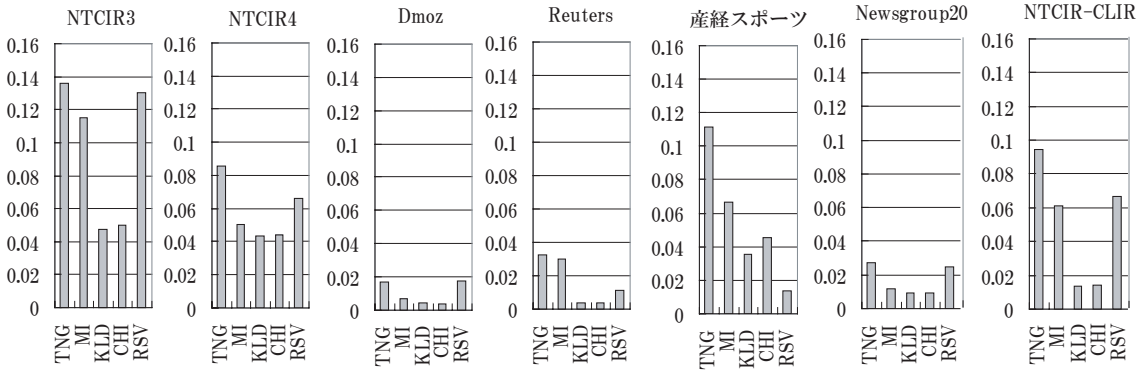


図 5 各手法によって得られた単語クラスタの $TP(t_i)$ 値による評価．7 つのデータを使用．
系列 A，B，C は，表 2 中の分類を参照．
Fig. 5 Evaluation of term clusters by sum of TP. We used seven different data.

Graph based on Building Construction Metaphor, *Proc. of IEEE ADL'98* (1998).

7) 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 (1999).

8) Takano, A., Niwa, Y., Nishioka, S., Iwayama, M., Hisamitsu, T., Imaichi, O. and Sakurai, H.: Associative Information Access Using DualNAVI, *Proc. of ICDL'00*, pp.285–289.

9) Robertson, S.E.: On term selection for query expansion, *Journal of Documentation*, Vol. 46, No.4, pp.359–364 (1990).

10) Fox, E. A., Das-Neves, F., Yu, X., Shen, R., Kim, S. and Fan, S.: Exploring the computing literature with visualization and stepping stones & pathways, *Commun. ACM*, Vol. 49, No.4, pp.52–58 (2006).

11) Yu, X., Das-Neves, F. and Fox, E.: Hard queries can be addressed with query splitting plus stepping stones and pathways, *Bulletin of the IEEE-CS Technical Committee on Data Engineering*, Vol.28, No.4, pp.29–38 (2005).

12) Sanderson, M. and Croft, B.: Deriving concept hierarchies from text, *Proc. of SIGIR 99*, pp.206–213 (1999).

13) Maki, M.: Findex: Search Result Categories Help Users when Document Ranking Fails, *Proc. of the SIGCHI conference on Human factors in computing systems*, pp.131–140 (2005).

14) Huo, H., Liu, J. and Feng, B.: Multinomial Approach and Multiple-Bernoulli Approach for Information Retrieval Based on Language Modeling., *FSKD (1)*, pp.580–583 (2005).

15) 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).

16) 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」，Vol.137, pp.269–276 (2005).

17) Baker, L. D. and McCallum, A. K.: Distributional clustering of words for text classification, *Proceedings of SIGIR-98*, pp.96–103 (1998).

18) I. S. Dhillon, S. M. and Kumar, R.: A divisive information-theoretic feature clustering algorithm for text classi cation, *Journal of Machine Learning Research(JMLR): Special Issue on Variable and Feature Selection*, pp. 1265–1287 (2003).

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

20) Toyoda, M., Kitsuregawa, M., Mano, H., Itoh, H. and Ogawa, Y.: University of Tokyo/RICOH at NTCIR-3 Web Retrieval Task, *Proc. of the 3rd NTCIR Workshop Meeting*, pp. 31–38 (2002).

21) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop (2003).

22) Eguchi, K., Oyama, K., Aizawa, A. and Ishikawa, H.: Overview of WEB Task at the Fourth NTCIR Workshop (2004).

23) Chen, K., Chen, H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S.H., Kishida, K., Eguchi, K. and Kim, H.: Overview of CLIR Task at the Third NTCIR Workshop (2003).

24) MeCab: http://mecab.sourceforge.jp/.

25) ipadic-2.5.1: http://chasen.naist.jp/stable/ipadic/.

26) Chakrabarti, S.: *Mining the Web:Discovering Knowledge from Hypertext Data*, Morgan-Kaufmann Publishers (2002).

(平成？年？月？日受付)
(平成？年？月？日採録)

若木　裕美

2002 年東京大学工学部電子情報工学科卒業．2004 年同大学大学院新領域創成科学研究科基盤情報学専攻修士課程修了．同年から東京大学大学院情報理工学系研究科電子情報学専攻博士課程に在籍．テキストマイニングの研究に従事．

正田　備也（正会員）

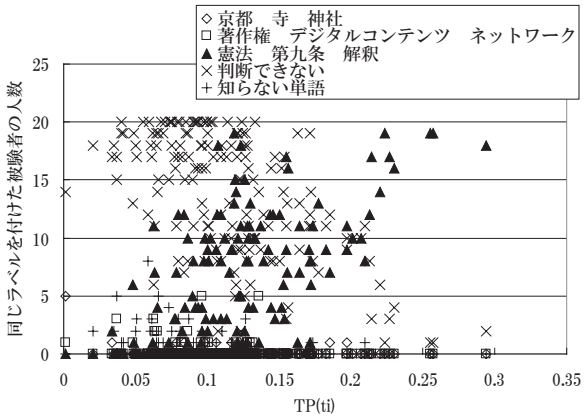
1970 年生．1995 年東京大学大学院理学系研究科情報科学専攻修士課程修了．1999 年東京大学大学院総合文化研究科広域科学専攻（科学史・科学哲学研究室）修士課程修了．1999 年富士写真光機 (株)（現フジノン (株)）入社．2004 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了．現在国立情報学研究所非常勤研究員．テキスト・マイニングの研究に従事．情報理工学博士．情報処理学会，日本現象学会，各会員．

高須　淳宏（正会員）

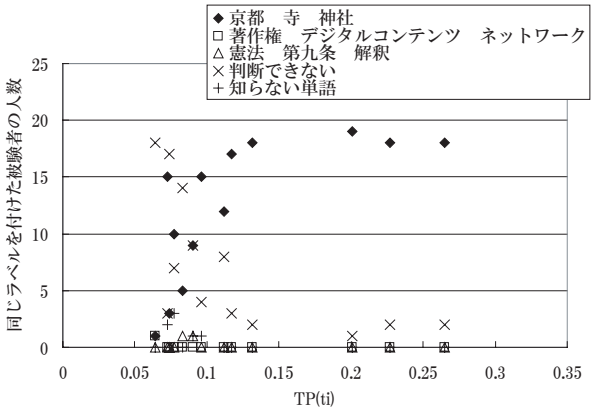
1984 年東京大学工学部航空学科卒業．1989 年同大学院工学系研究科博士課程修了．工学博士．同年，学術情報センター研究開発部助手．同センター助教授．国立情報学研究所助教授を経て 2003 年より同研究所教授．データ工学，特にデータ解析と解析モデルの学習の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，ACM，IEEE 各会員．

安達　　淳（正会員）

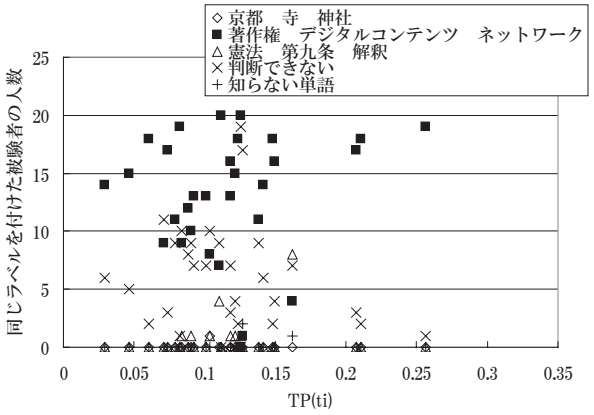
1981 年東京大学大学院工学系研究科博士課程修了．工学博士．東京大学大型計算機センター助手，文部省学術情報センター助教授、教授等を経て現在国立情報学研究所教授．東京大学大学院情報理工学研究科教授を併任．データベースシステム，テキストマイニング、情報検索，電子図書館システム等の研究開発に従事．電子情報通信学会，情報処理学会，日本データベース学会，IEEE，ACM 各会員



a)「憲法，第九条，解釈」



b)「京都，寺，神社」



c)「著作権，デジタルコンテンツ，ネットワーク」

図 6 NTCIR3 の検索質問「憲法，第九条，解釈」,「京都，寺，神社」,「著作権，デジタルコンテンツ，ネットワーク」のそれぞれの正解集合を混ぜたデータから抽出した単語それぞれについて，$TP(t_i)$ の値と被験者の主観評価の関係を表す散布図．

Fig. 6 Scattergrams to show correlation between TP evaluation and subjective evaluation of terms obtained from NTCIR3's data and queries.