トピック指向単語クラスタリングを用いた 複数トピックの包括的提示による検索支援

若木　裕美†　　正田　備也††　　高須　淳宏††　　安達　　淳††

† 東京大学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　本研究では，検索結果に含まれる複数のトピックを分離し，各トピックを表すような単語クラスタを提示することによって，ユーザによる検索質問の改善を支援することを目的とする．また，それらの単語は検索質問拡張として使うことができる．さらに，提案手法の有効性を確認するために主観的評価実験も行った．実験の結果から，提案手法は検索質問拡張と検索語に関連する話題の発見に有効であると考えられる．

キーワード　情報検索, 知識発見, Web とインターネット

Query Refinement based on Comprehensive Representation of Multiple Topics using Topic-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　We aim to present multiple topics contained in the search results by showing term clusters each of which corresponds to one of the topics. These term clusters can be used as query terms for query expansion. Our experimental results by human subjective evaluation show that our method is suitable for query refinement and knowledge discovery related to the original query.

Key words　Information Retrieval, Knowledge Discovery, Web

1. はじめに

現在の検索エンジンでは，ユーザによって入力された検索語に関連する文書の中で，より検索語と関係が深いと思われる文書が一次元的にランキングされる．ユーザ側もこの検索エンジンの特性に合わせて，既に必要とするものがはっきりと分かっているときにブックマーク的に使うことが多い．しかし Web 上には様々な内容の文書が存在し，検索結果としてトップページに表示される中には多様なトピックが混在している．そしてユーザが曖昧な検索語を入力すれば，多くのトピックが混在した状態でランキングがなされ，自分の必要とするトピックを見つけ出すことが難しい．例えば，あまりよく知らない言葉について調べたいと思ったときに，従来の検索エンジンを使っても幅広く調べることができない．様々なページを閲覧して，トピックが絞りこめるような単語をユーザ自身が発見して追加する必要がある．

このような背景をふまえ近年では，Clusty(注1)のような文書クラスタリング型の検索エンジンが幾つか登場している．Clusty ではメタサーチを行って複数の検索エンジンの結果を文書クラスタリングして，各クラスタに名前を付ける．文書クラスタとその名前をユーザに提示することにより，多くの検索結果を整理することを目的としている．また，Google Suggest のようにログベースで検索質問を提示する方法や，Google の関連検索として検索語に追加を促す単語が表示されるようになってきた．いずれの方法でも多くの人が入力した単語や多くの文書が関わる一般的な単語が表示されやすい．

そこで我々は，多様な内容を含む検索結果の中から，含まれ

(注1)：Clusty 日本語サイト http://clusty.jp/,
Clusty 英語サイト http://clusty.com/

る複数のトピックを分けるのに効果的な検索語を提示し，検索語の曖昧性を解消する手法を提案する．また，各単語を使って検索を行うと新しい情報が得られるような単語であることを期待している．一般的に多義性解消で求められるのは辞書的な複数の意味に分けることであるが，検索語は１～2 語であることが多く [14]，固有名詞は複数の概念や対象を示しうる．このような多義性の解消は，個々のトピックを分離することで解決できると考えている．また，検索語の示すものが一意に決まる場合においても，同一のものを異なる視点から見ることで異なった話題が考えられる．これもまた，個々のトピックを分離することで解決できると考えられる．

本研究では，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．本稿では，実際の検索エンジンの検索結果の上位を用いて本手法を適用した単語クラスタを生成する．また，主観的評価実験を通じてこの単語クラスタを評価する．本実験では，検索結果に含まれるトピックを分けて示すことで検索語自身の多義性の発見に役立つだけでなく，検索語に関連する新しい話題の発見につながることを評価実験から明らかにした．

2. 関 連 研 究

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が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．

まず，検索結果の中に混在する複数のトピックを分けて提示する方法として，文書を分類あるいはクラスタリングする方法があげられる [3] [9] [18] [23]．Scatter/Gather [3] [9] では文書をクラスタリングして，各分類にフレーズでラベルをつけている．しかし，これでは名前が分かりにくいという問題がある．そこで，近年では，Clusty や vivisimo(注2)，Grokker(注3)，WAKANO(注4)といった，検索結果を整理することを目的とした検索エンジンが数多く出てきている．また検索結果を分類することが検索支援として有効であることを示した研究として Findex [15] がある．Findex システムでは，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．そして，そのシステムを被験者に使ってもらい，使用のログとアンケートによって検索結果が分けられていると効果的な場合について議論している．特に，検索語が曖昧で一般的な単語であるときは，ユーザ自身が元々複数のトピックを閲覧したいことが多く，複数のトピックについて閲覧を行っていたという結果が得られている．

いずれも検索結果の整理を目的としているが，本研究では単語のクラスタを作ることによってトピックを提示することを目的としており，文書のクラスタリングは行わない．このため，選択された単語を使って検索拡張を行って結果を得る．つまり，トピックに関連する文書を最初の検索結果の中で見つけるのではなく，新たな検索によって取ってくるという方法である．

また近年，Faceted Search が注目されている．検索結果をファセットと呼ばれる異なった視点から見て分類することで，文書のクラスタリングよりも効果的にユーザが内容を閲覧できるというものである [7]．ファセットは直交する異なるカテゴリの属性を指す [1] [4] [8] [17]．例えば，“アクセサリー” のショッピングサイトでいうと，ファセットとして “素材” や “種類” がある．そして，“素材” については “ゴールド” や “シルバー”，“種類” については “リング” や “ネックレス” がそれぞれのファセットに対する値となる．ファセットという考え方自体は S. R. Ranganathan が 1930 年代に考えたものであるが，Faceted Search として情報検索に適用することが注目されている(注5)．Faceted Search としてはメタデータが付いていたり，事前に複数の独立したファセットに整理された文書を扱うことが多く [20]，通常の検索結果に対して自動的にファセットを与えるのは難しい．我々の提案する手法では，自動的に検索結果の中に含まれるファセットを提示することが期待される．

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) [19] 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 [6] [22]．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．本研究では，システムが提示する単語を使って，検索結果の中に含まれる複数のトピックのそれぞれに検索を絞り込めることを期待している．

3. TNG を用いた複数トピックの提示システム

3.1 TNG:トピックを強く表す単語の抽出とそのクラスタリング

3.1.1 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

(注2)：http://vivisimo.com/

(注3)：http://www.grokker.com/

(注4)：http://www.wakano.co.jp/

(注5)：第一回目の Workshop on Faceted Search が SIGIR'06 において開催された (http://facetedsearch.googlepages.com/)．

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 [25](注6)．

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．

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

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．

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

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，

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

で定義したところを, 以下のような式によって Dirichlet smoothing [10] を行った．

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

「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 [24] [21]．ただし, [24], [21] 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．

3.1.2 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．

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

（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 [2] [11] を用いる（Table. 1）．

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 による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，

(注6)：定式化とその有効性については [25] でも検証している．本稿では，特に実際の検索結果に対して適用した．

表 1　Distributional Clustering アルゴリズム.

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.

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

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，

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

と定義する．そして，閾値 $X_{threshold} = 0.01$ より類似度の高いクラスタペアをすべてマージした．このため，この単語クラスタリングはクラスタ数が自由に決まる．また本研究では，$M = 20$，単語は上位の 200 語とした.

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 [13]．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.

3.2　システムの構成

本システムでは，検索語が入力されると Google の検索結果の上位およそ 500 件の URL を取得する．そして次に，URL から実際の文書をダウンロードする．ダウンロードした文書中で HTML に関係するタグを除去し，テキスト情報だけを得る．各文書の長さは文書ごとによって大きく異なるため，1 つの文書に出現する単語の共起情報を使う本研究の手法では，全体の文書を使うのは効果的でない．そこで，各検索結果から同一の長さのスニペット（要約情報）を生成する．詳しい手順は後述する.（3.2.2 節を参照）．この要約情報を基に TNG の計算（3.1.1 節参照）を行って，単語クラスタを生成（3.1.2 節参照）する．全体の流れを図 1 に示した.

3.2.1　Google の検索結果からの文書のダウンロード

各検索語に対する Google の検索結果上位の 500 件程度について URL を取得し，その URL を基に実際の文書をダウンロードする．そして，各文書から HTML タグを除去した.

3.2.2　要 約 生 成

各文書ごとで出現する単語数や文書長，単語の種類数が大きく異なるため，本手法のように各文書中における共起を測る場合には，頻出話題が大きく取り上げられすぎる傾向がある(注7).

(注7)：予備実験として，2 つの実験を行った．1つは Google の出力するスニペットを1つの文書としてみなして実験，もう 1 つは URL からダウンロードできる実際の文書全体を用いた実験である．前者の方が後者よりも結果が良かった．これは，文書全体とすると多くの話題が含まれすぎること，文書長が異なると出現する単語数や単語の種類に差があり小さいトピックが負けてしまうなどの原因が考えられる．そこで，これらの原因を解決するような一定長の要約を利用することとした.

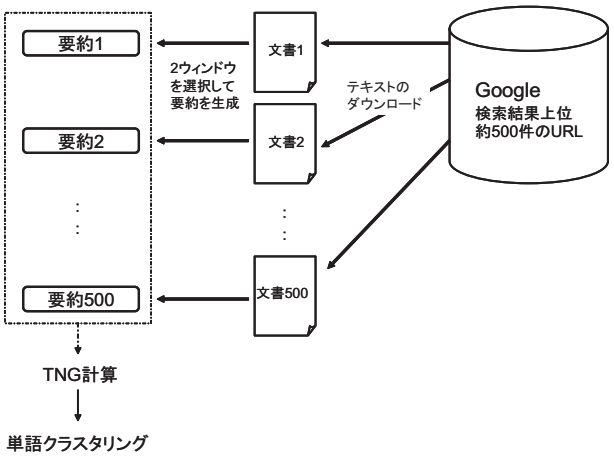


図 1　本手法によるシステムの処理手順の概要.

そこで，検索語を含む窓幅 30 語が 2 セットからなる要約を各文書に対して用意する．つまり，次のような 3 つの情報を基に重みの最も高い 2 つの窓を，各文書ごとに生成しこれを各文書とみなす.

- 検索語の出現回数
- 検索語の種類数（複数語からなる検索語のときに有効）
- 名詞の個数

先頭の 1Mbyte 中でかつ 10 個目の検索語までを対象として，上記の窓を計算した．ストップワードとしては，一般的に HTML 文書に現れやすい単語，英語でストップワードとして使われやすいもの，日本語で頻出する単語を除いた．MeCab [16] を用いて形態素解析を行った．また，辞書は ipadic-2.5.1 [12] である.

3.2.3　提案手法の適用

要約文書を用いて 3.1.1 節で提案する TNG の計算を行う(注8)．そして TNG の高い順にソートし，高い方から 200 語を対象として，3.1.2 節の単語クラスタを行う．ここで，各要約文書中に現れる単語同士は共起したとみなす(注9)．また，最初においておくクラスタ数は $M = 20$ とした．クラスタ数は自動的に決まるが，人間による主観的評価実験の際には，単語クラスタ中にある単語の TNG の高い順に 10 個のクラスタだけを利用する．また，各単語クラスタ中にある単語は，DF の多い順にソートし上位の 5 語だけを利用し，5 語より小さいクラスタの場合には全部が提示されることになる（図 2 参照）.

3.3　システムの出力結果について

本システムの出力した単語クラスタの上位 10 個の例を表 2 表 3 に載せた．表 2 は検索語 “ジャガー”，表 3 は検索語 “ニンテンドー DS” の結果である．例えば，表 2 中では，クラスタ 1 が車の種類の列挙となっていて車のジャガーが示唆される．クラスタ 2 はプロレスラーのジャガー横田に関する話題が示唆される．クラスタ 3 は車のジャガーの xkr シリーズ，クラスタ 4 はネコ科のジャガー，クラスタ 5 はジャガーミシン，クラスタ 6 は車のジャガーの中古車，クラスタ 8 は腕時計のジャガールクルト，クラスタ 10 は最新のジャガーの車種が示唆される.

(注8)：document frequency の多い 500 語だけを対象として計算を行う.
(注9)：単語の共起は，1 回しか共起しないときはノイズとみなして無視した

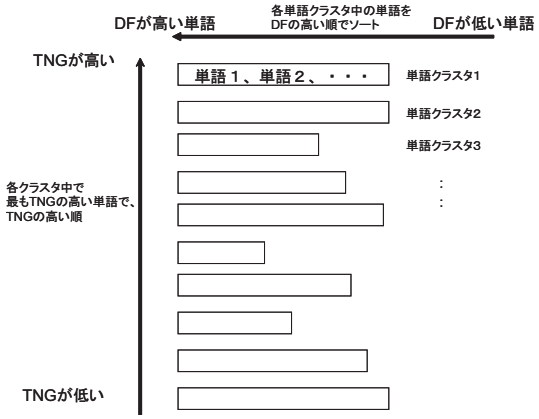


図 2　提示した単語クラスタと単語クラスタ内の単語のソート方法

表 2　検索語が “ジャガー” のとき.

クラスタ番号	単語クラスタ
1	フォード トヨタ bmw ポルシェ アウディ
2	横田 出産 女子 木下 プロレスラー
3	new 注文 搭載 開始 xkr
4	動物 ネコ アメリカ 妊娠 食肉
5	ミシン 説明 net ロック 電子
6	中古 輸入 ディーラー 正規 認定
7	買取 記入 高額 完了 必須
8	ルクルト 腕時計 lecoultre jaeger
9	内容 著者 投票 書名
10	車種 最新 使用

表 3　検索語が “ニンテンドー DS” のとき.

クラスタ番号	単語クラスタ
1	発表 ドラゴンクエスト ドラクエ ドラゴン 新作
2	アクション シミュレーション ロールプレイング アドベンチャー パズル
3	プレイステーション xbox キューブ ゲームボーイアドバンス ボーイ
4	常識 監修 モンスター 検定 ダイヤモンド
5	任天堂 nintendo 関連 amazon クリスタルホワイト
6	無線 lan usb コネクション 接続
7	開発 opera 共同 software
8	拡張 カートリッジ メモリー
9	ブラック ホワイト クリスタル ジェット
10	トピックス 経済 ウェブ 地域 floor

4.　複数トピック提示システムの主観的評価実験

4.1　比較対象

Web の検索結果の整理を行うものとして Clusty(注10) を比較対象とした．Clusty に検索語を入力すると左側に出力されるクラスタのラベルと TNG による単語クラスタの比較を行った．通常 Clusty では上位の 10 個だけがデフォルト出力されるため，いずれの結果も上位の 10 個のクラスタのラベルを用いた．

4.2　実験参加者

20〜30 代の男性 8 名に参加してもらって実験を行った．検索を使うようになってから 6〜10 年の経験を全員が持つ．また，専門分野の内訳は，情報検索・テキストマイニングが 3 名，通信ネットワークが 2 名，CG・画像処理が 2 名，ポートフォリオ・マネージメントが 1 名であった．普段使用する検索エンジン（複数回答可）としては，Google が 8 名，Yahoo!JAPAN が 2 名，MSN が 2 名，goo が 1 名であった．普段入力する検索語の数は，3 語以内が 6 名，5 語以内が 1 名，10 語以内が 1 名であった．普段閲覧するページ数の上限が，10 件以内が 1 名，30 件以内が 4 名，100 件以内が 1 名，200 件以内が 1 名，300 件以内が 1 名であった．

(注10)：http://clusty.jp/

表 4　実験に使用した検索語 20 語の一覧とその出典．(Yahoo!Japan2006 検索ワードランキングを “Yahoo” と書く．)

検索語	出典
飛行機	一般的な単語
北朝鮮	ニュース
アムステルダム	一般的な単語（地名）
無免許	ニュース
クラスタリング	専門用語
野村	意味が複数ありそうな一般的な単語
情報	意味が複数ありそうな一般的な単語
アップル	多義語として有名
ジャガー	多義語として有名
KAT-TUN	Yahoo（男性ランキング 1 位）
ニンテンドー DS	Yahoo（商品検索数ランキング 1 位）
mixi	Yahoo（総合ランキング 1 位）
あいのり	Yahoo（テレビ番組ランキング 1 位）
涼宮ハルヒの憂鬱	Yahoo（検索トレンド分析）
ワンセグ	Yahoo（検索トレンド分析）
成分分析	Yahoo（検索トレンド分析）
DEATH NOTE	Yahoo（漫画アニメランキング 1 位）
哺乳類 絶滅	NTCIR3 Web 検索タスク
北欧神話 世界樹	NTCIR3 Web 検索タスク
レオナルド・ダ・ヴィンチ	映画 “ダ・ヴィンチコード” より

4.3　検　索　語

表 4 にある 20 語の検索について実験を行った．表 4 には使用した検索語とその出典を載せた．このうち，8 個は Yahoo!Japan による 2006 検索ワードランキング(注11)(注12) からのものである．総合ランキング 1 位の “mixi”，男性ランキング 1 位の “KAT-TUN”，テレビ番組名ランキング 1 位の “あいのり”，漫画アニメ名 1 位の “DEATH NOTE”，商品検索数ランキング 1 位の “ニンテンドー DS” などを利用した．また，それ以外の単語として，多義語として有名な “ジャガー”，“アップル” や，NTCIR3 の Web 検索タスク [5] で用いられた単語，一般的に認知度の高い単語などを利用した．

4.4　実験前アンケートの結果

実験を始める前に 20 個の検索語について表 5 の q1〜q4 のアンケートを行った．ただし，q1，q3，q4 については，Y（はい）または N（いいえ）で答えてもらった．q2 については，Y(はい) または N(いいえ) または M(普通) で答えてもらった．q1〜

(注11)：http://picks.dir.yahoo.co.jp/new/review2006/general.html
(注12)：集計期間：2006 年 1 月 1 日〜2006 年 11 月 5 日

表 5　q1〜q4 のアンケートの結果．各アンケートに対して，8 人の被験者が 20 語の検索語に Y（“はい”）または N（“いいえ”）または M（“普通”）と答えた数．(() 内は 1 人あたりの平均語数．)

問	Y	M	N
q1: 聞いたことがあるか？	156 (19.5)	-	4 (0.5)
q2: よく知っているか？	36 (4.5)	96 (12.0)	28 (3.5)
q3: 多義語だと思うか？	52 (6.5)	-	108 (13.5)
q4: 複数の話題があると思うか？	119 (14.9)	-	41 (5.1)

q4 のいずれも，全員が全ての検索語に Y と答えた場合，総計 8(人)×20(検索語)= 160 となる数である．表 5 を見ると，ほぼ全員がすべての検索語について聞いたことがあり，またある程度知っている単語がほとんどである．このことから，すべての検索語について前提知識として多少の知識があるといえる．また，多義語だと思うものは 20 語の検索語のうち 6.5 語，複数の話題があると思うものが 15 語程度である．

4.5 実験方法

各検索語について，Clusty の出力した単語クラスタ 10 個と TNG を使った提案システムの出力した単語クラスタ 10 個を被験者に提示する．左側に Clusty の出力した単語クラスタを並べ，右側に TNG の出力した単語クラスタを並べる（図 3 参照）．そして，各検索語について提示されている 20 個の単語または単語クラスタ（左側 10 個が Clusty，右側 10 個が TNG の出力結果）を検索語にそれぞれ追加して検索に使ってもらう．この作業を表 4 中の 20 個の検索語すべてについて行い，下記の Q1〜Q3 のアンケートに各検索語ごとに答えてもらう．ただし被験者は，どの単語クラスタのセットがいずれのシステムから出力された単語であるかを知らない．

各検索語ごとに，Clusty による単語クラスタ群と TNG による単語クラスタ群のそれぞれについて，

- Q1: 複数の意味が発見できたか？
- Q2: 思いがけない話題が発見できたか？
- Q3: 検索語についての知識が増えたか？

に Y(はい) または N(いいえ) で答えてもらった．全員が全ての検索語に Y と答えた場合，総計 8(人)×20(検索語)= 160 となる数である．

また，検索実験の後，検索語 20 個の全体を通じた感想として，下記の Q4〜Q5 のアンケートに答えてもらう．

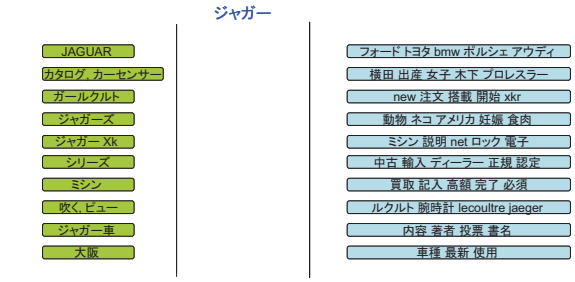
- Q4: 検索に使いやすいのはどちらか？
- Q5: 面白い話題が発見できたのはどちらか？

のそれぞれに緑（Clusty）と青（TNG) で答えてもらった．もし全員が緑と答えたら 8（人）となる数である．

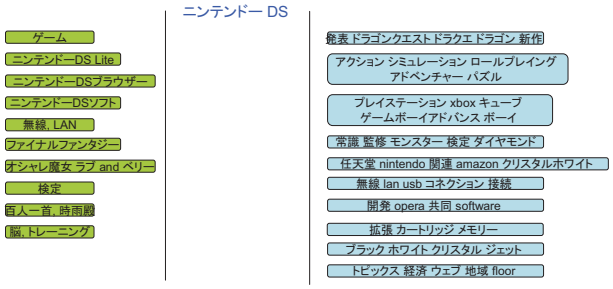
4.6 実験結果

検索語と共に提示した単語を検索語に追加する単語として使った実験の後のアンケート結果は，表 6 表 7 である．表 6 では，Clusty によって出力された単語と TNG によって出力された単語のそれぞれについて，各検索語ごとに，質問に対し Y(はい) と答えた人数の総和を記載した．全員が全ての検索語に Y と答えた場合，総計 8(人)×20(検索語)= 160 となる数である．

表 6 中 Q1 の結果から Clusty，TNG のいずれの出力結果か



a) 検索語 “ジャガー” のとき



b) 検索語 “ニンテンドー DS” のとき

図 3　実際に被験者に提示した Clusty の出力する単語（左側）と TNG による単語クラスタ（右側）

らも同じくらい複数の意味が発見できたと考えられる．実験前のアンケート結果で 20 個の検索語のうち多義語だと思うと答えられたものが 6.5 語であったのに対し，実験後のアンケートでは 10.5 語と増えていた．このことから，関連するトピックを複数個提示されることで，多義性に気が付けるという利点があることが分かる．また，思いがけない話題を発見できた数 (表 6 中 Q2) は，Clusty の出力した単語に比べて TNG による単語の方が多かった．更に，検索語についての知識が増えた数 (表 6 中 Q3) も，Clusty の出力した単語に比べて TNG による単語の方が多かった．実験前アンケートでは複数の話題が思い浮かんだものが 20 個の検索語のうち 15 語程度あったが，それでも複数のトピックを提示されることでさらに思いがけない話題を Clusty よりも発見できていたことが分かる．また，検索を通して，検索語に対する知識が増えたと答える人が多かったのは，検索語に関する個別のトピックが絞り込んで得られたためであると考えられる．

検索に使いやすいのは Clusty であると大多数の人が答えた (表 7 中 Q4)．単語クラスタという形になっていると閲覧がしにくい点が原因と考えられる．また，Clusty の方が広く浅い単語が表示されるため，あまり知識がない場合には分かりやすいと考えられる．これに対して，面白い話題が発見できたものとしては TNG であると大多数の人が答えた (表 7 中 Q5)．これは，実際に使ってみた結果として，個別のトピックに関する深い話題が発見できているといえる．

ここで，Clusty と TNG のそれぞれのシステムで提示される単語数は同一ではない．しかし，本実験は個々の単語による知識の増加を望むものではなく，各単語クラスタを元の検索語に追加して得られた検索結果を見て，新たな知識が得られたか，面白い話題が含まれていたかを判断するものである．提示され

表 6　Q1〜Q3 のアンケートの結果．各アンケートについて Clusty が出力した単語と TNG を使ったシステムが出力した単語のそれぞれに対して，8 人の被験者が 20 語の検索語に Y（“はい”）と答えた数．(() 内は 1 人あたりの平均語数である．)

問	Clusty	TNG
Q1: 複数の意味が発見できたか？	84 (10.5)	85 (10.5)
Q2: 思いがけない話題が発見できたか？	73 (9.1)	104 (13.0)
Q3: 検索語についての知識が増えたか？	90 (11.3)	123 (15.4)

表 7　Q4〜Q5 のアンケートの結果

問	緑（Clusty）と答えた人数	青（TNG）と答えた人数
Q4: 検索に使いやすいのは？	7	1
Q5: 面白い話題が発見できたのは？	1	6

る単語数が多いとき，多くの情報を含む可能性もあるが，クラスタが適切に生成されていなければ検索結果の絞込みがうまくいかないことになり，単語数によって有利不利となるような実験ではない．

5. 考　　察

5.1　検索語の曖昧性の解消

検索語による曖昧性は，幾つかの要因が考えられる．第一に，検索語が少ないために複数の解釈が可能となり，検索語が複数の概念や対象を示しうるために多義性が見られる場合がある．第二に，検索語が示すひとつの概念や対象を様々な視点から見ることが可能であるというファセットの問題の場合がある．

（1）　多義性の問題に対して

検索語に関連するトピックに分け，本手法のように単語クラスタで表すことによって，多義的な検索語の曖昧性を分けることができると考える（図 4）．当初多義的であると思っていなかったものであっても，このように単語クラスタで表すことによって複数の意味の発見に役立つ．そして，いずれかの意味の内容だけを検索するのに有効な単語を利用することが容易になる．例えば，検索語 “ジャガー” の例では複数の意味があることが各単語クラスタにおいてどのような単語群が束ねられているかを見比べることで推測できる（図 6）．

（2）　ファセットの問題に対して

また，検索語が示すひとつの概念や対象がもつ様々な側面を単語クラスタによってファセットとして表すことができる．ファセット自体を言葉で表すことができなくてもファセットの値をまとめることで，ファセットが何であるかを想像することが可能である（図 5）．ファセットは階層的であると想定されるが [7]，本手法では階層構造については表すことができない．しかし，Clusty と異なり単語クラスタを積極的に生成することで，ファセットの構成要素を示唆することができると期待できる．検索語 “ニンテンドー DS” の例では，同じ階層でも異なるファセットがあることが単語から推測できる（図 7）．Clusty では，提示された単語（または単語クラスタ）の中には同一ファセットの値になるものがばらばらになっていることがあるため，検索語に対してどういった役割を果たすのかが分かりにくい．

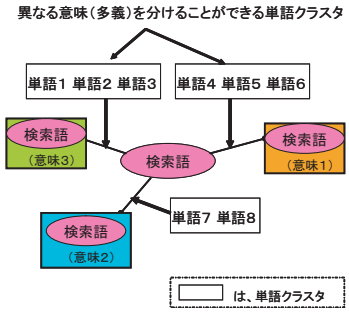


図 4　多義語を分けるのに役立つ単語セットの場合

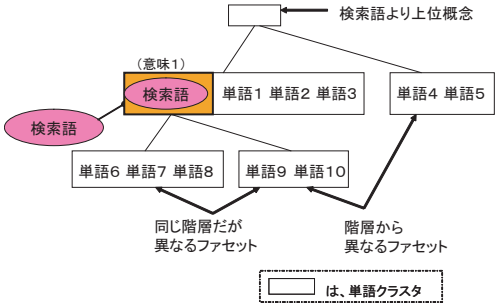


図 5　ファセットを表す単語セットの場合

5.2　検索語に関する特徴的な話題の発見

このように “異なる意味を見出すような単語群” あるいは “異なるファセットになる単語群” を明示的に表すことで，検索にとってより効率的な絞込みが可能であると考えられる．特に，被験者のアンケート結果では，面白い話題が発見できたのは TNG のほうであると大多数が答えた（表 7 中 Q5）．これは，特徴的な話題を発見する機能が優れているからであるといえる．

また，実験後に検索実験についての感想として次のような質問 1，2 について訊ねた．左側（緑）をシステム 1 による結果，右側（青）をシステム 2 による結果とした場合に，

- 質問 1：どのようなときにシステム 1 やシステム 2 を使うと Google より効果があると思いますか？
- 質問 2：またその場合にシステム 1 とシステム 2 のどちらがいいと思いますか？

この二つの質問の結果，質問 1 の答えとしては，次のようなものがあった．検索語に複数の意味がある場合，検索語に関連して検索をする場合，対象となる言葉について深く調べる際に何を元に探せばいいか分からない場合，対象となる言葉について自分の着目している視点以外からのアプローチを得たいとき，などの場合に効果的であると答えた．また，質問 2 については異なる場面でそれぞれのシステム共に使いやすい面があるという意見が多かった．そして，具体的に対比させた感想が多くあったので参考として表 8 に載せた．また別の意見として，Clusty では閲覧がしやすいが，曖昧さが残ってしまう場合があるという意見もあった．

6. ま　と　め

本研究では，検索質問の改善のために，検索結果の中に含まれる複数のトピックを分離し各トピックを表すような単語クラスタを提示する手法を提案した．また，これらの単語クラスタを使って検索質問拡張を行う主観的評価実験を行った．実験の

表 8　実験後のアンケートの中で Clusty と TNG を対比させていた感想

被験者 ID	Clusty	TNG
被験者 1	広い検索	深い検索
被験者 2	語句の定義を調べる	その語句に関連した話題を調べる
被験者 3	分けたい話題や意味がある程度しっかりしている場合	分けたい話題や意味に偏りが大きい場合
被験者 4	浅い知識の検索	面白い話題探しの検索

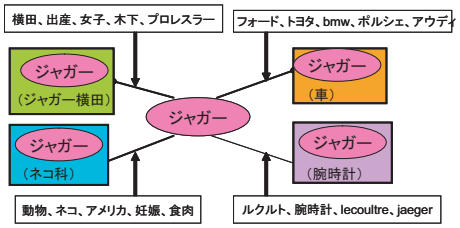


図 6　検索語 “ジャガー” についての単語クラスタによる多義語の表現

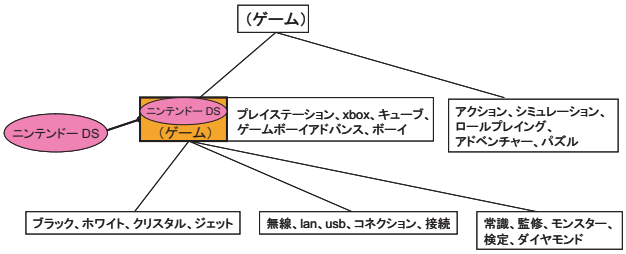


図 7　検索語 “ニンテンドー DS” についての単語クラスタによるファセットの表現

結果から，提案手法は検索語に関連する話題の発見に有効であるといえる．また，特定の話題について効率的に検索を絞り込める単語クラスタとなっていると考えられる．今後は，単語クラスタを見せるのではなく，単語クラスタ中の軸となる単語だけを選択して提示するなどの改善を行う予定である．

検索結果を表示する際に文書のランキングを提示する場合，次のような二つの曖昧性が考えられる．検索語が複数の概念や対象を示すという多義性の問題と，検索語が示すひとつの概念や対象を様々な視点から捉えることが出来るというファセットの複数性の問題である．検索結果に含まれる複数のトピックを分けて見せ，また，各トピックを複数の単語セットによって表すことによって，この二つの問題が解決できると期待している．

文　献

[1] H. P. Adkisson. Use of faceted classification, 2005. http://www.webdesignpractices.com/navigation/facets.html.

[2] L. Douglas Baker and Andrew K. McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR-98*, pp. 96–103, 1998.

[3] D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. of SIGIR'92*, pp. 318–329, 1992.

[4] W. Denton. How to make a faceted classification and put it on the web, 2003. http://www.miskatonic.org/library/facet-web-howto.html.

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

[6] E. A. Fox, F. Das-Neves, X. Yu, R. Shen, S. Kim, and S. Fan. Exploring the computing literature with visualization and stepping stones & pathways. *Communications of the ACM*, Vol. 49, No. 4, pp. 52–58, 2006.

[7] M. A. Hearst. Clustering versus faceted categories for information exploration. *Communications of the ACM*, Vol. 49, pp. 59–61, 2006.

[8] M. A. Hearst, P. Smalley, and C. Chandler. Faceted metadata for information architecture and search. In *CHI 2006 Course*, 2006.

[9] M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proc. of SIGIR'96*, pp. 76–84, 1996.

[10] H. Huo, J. Liu, and B. Feng. Multinomial approach and multiple-bernoulli approach for information retrieval based on language modeling. In *FSKD (1)*, pp. 580–583, 2005.

[11] S. Mallela I. S. Dhillon and R. Kumar. 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.

[12] ipadic-2.5.1. http://chasen.naist.jp/stable/ipadic/.

[13] 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.

[14] B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Searchers, the subjects they search, and sufficiency: A study of a large sample of excite searches. In *1998 World Conference on the WWW and Internet*, 1998.

[15] M. Maki. Findex: Search result categories help users when document ranking fails. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pp. 131–140, 2005.

[16] MeCab. http://mecab.sourceforge.jp/.

[17] P. Morville. *Ambient Findability*. O'Reilly Media, 2005.

[18] W. Pratte and L. Fagan. The usefulness of dynamically categorization search results. *Journal of the American Medical Informatics Association*, Vol. 7, pp. 605–617, 2000.

[19] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, Vol. 46, No. 4, pp. 359–364, 1990.

[20] D. Tunkelang. Dynamic category sets: An approach for faceted search. In *Proc. of the ACM SIGIR 2006 Workshop on Faceted Search*, 2006.

[21] H. Wakaki, T. Masada, A. Takasu, and J. Adachi. A new measure for query disambiguation using term co-occurrences. In *Proc. of 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pp. 904–911, 2006.

[22] X. Yu, F. Das-Neves, and E.A. Fox. 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.

[23] O. Zaimir and O. Etzioni. Grouper: A dynamic clustering interface to Web search results. In *Proc. of WWW8*, pp. 1361–1374, 1999.

[24] 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.

[25] 若木裕美, 正田備也, 高須淳宏, 安達淳. 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. 情報処理学会論文誌データベース (TOD), Vol. 47, No. TOD32(SIG19), pp. 72–85, 2006.