Vol.2010-DBS-151 No.2
2010/11/12

文書クラスタリングによる話題の絞込みを利用した先読み検索

藤田遼治[†1]　太田学[†1]

我々は，ユーザの検索を観察することで，ユーザの検索意図を推測し，ユーザの代わりに検索を行う先読み検索を提案してきた．本稿では，特に絞り込み検索を対象とした先読み検索の精度改善について報告する．本研究では，検索結果の変化を利用した文書クラスタリングにより得られた話題集合から，ユーザの検索目的に沿った話題を抽出する．さらに，抽出した話題中の語句を用いて検索質問を拡張することで先読み検索を行う．実装したプロトタイプシステムを用いた実験により，話題の抽出精度および先読み検索の精度改善を定量的に評価した．

A Prediction Search Using Topic Distillation by Document Clustering

Ryoji FUJITA[†1] and Manabu OHTA[†1]

We have proposed a prediction search which searches for an automatically generated query on behalf of a user by observing his/her search. In this paper, we present a method to improve precision of the prediction search for narrowing search results. First, we extract topics from search results by clustering search results based on their change. Second, we select a topic which best matches a user's search intention. Then, we perform a prediction search for an expanded query including the terms extracted from the selected topic. By the experiment with our implemented prototype system, we quantitatively evaluated accuracy of extracted topics and the improvement in precision of the prediction search.

†1 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

1. はじめに

Google[⋆1]に代表されるサーチエンジンは Web 上の情報とユーザとの接点として重要な役割を果たしている．しかし，Web 上のコンテンツは膨大であり，一度検索しただけではユーザが求める情報をいつも得られるとは限らない．そのためユーザは通常，求める情報が得られるまで検索質問を修正し再検索を繰り返す．一度の検索時間は短いが，適切な検索質問が思いつかなければ必要な情報を得るまでに思わぬ時間がかかってしまうことも少なくない．

我々は，ユーザの連続した検索に着目し，ユーザの検索の先を見越し，検索意図に沿った検索質問を自動生成して検索する先読み検索を提案した[1)–3)]．そのなかで検索質問の変化をキーワードの追加，削除，変更の 3 種類に分類し，検索意図の推測を試みた．

本稿では特に，この中の追加変化を対象とし，検索精度の向上を目指す．追加変化においてユーザは，検索結果を特定の話題に絞り込もうとしていることが推測される．そこで，検索結果文書をクラスタリングすることで，ユーザの絞り込もうとしている話題を特定し，特定した話題を利用して先読み検索を行う．具体的には，検索結果文書集合をクラスタリングすることで，検索結果文書を複数の話題クラスタに分類する．そして，検索結果中に現れる特徴語の出現頻度の変化に基づいて，ユーザの検索意図に合致した話題クラスタを抽出する．その話題クラスタを構成する検索結果文書中の特徴語を利用して，検索質問を自動生成して先読み検索を行う．

本稿ではまず，2 章で関連研究について述べる．3 章で提案手法を説明し，4 章でプロトタイプシステムの実装について述べる．さらに，5 章で実験を行い，6 章でまとめと今後の課題について述べる．

2. 関連研究

本章では，本研究に関連する主な研究として，2.1 節で検索結果からの検索語候補の取得に関する研究，2.2 節で検索過程を用いた検索語想起支援に関する研究，2.3 節で文書クラスタリングを用いた話題の抽出に関する研究について述べる．

2.1 検索結果からの検索語候補の取得

検索結果から後の検索に役立つであろう検索語候補を取得し，ユーザに推薦する研究[4)]がある．検索結果から語の出現頻度に基づくスコアを用いて検索語候補を取得し，ユーザに提

⋆1 http://www.google.co.jp/

© 2010 Information Processing Society of Japan

示する．ユーザは提示された語から任意の語を選択することにより，検索質問を拡張することができる．

また，単語の出現する文同士の距離から関連単語を抽出し，検索質問拡張に利用する研究[5]もある．検索結果中の単語の出現頻度，および検索語が含まれる文と単語が含まれる文との距離からスコアを計算し，スコアが上位の単語を検索質問に追加することで検索質問拡張を行う．本研究でも検索結果を用いて検索質問を拡張するが，文書クラスタリングや，検索結果の変化を利用して検索質問を拡張する点等が異なる．

2.2 検索過程を利用した検索語想起支援

検索結果だけでなく，検索の全体像をデータフロー図として表示することで，検索過程の処理をユーザに分かりやすく提示する試み[6]も行われている．データフロー図に基づいた視覚的なインタフェースを用いて検索条件や検索結果を提示することにより，ユーザが検索過程や処理の結果を理解しやすく，またデータフロー図を直接操作することで，検索条件の修正等が簡単に行えるようになっている．中岡ら[7]は取得した検索語候補をグラフで表示することにより，視覚的に検索を進める手法を提案した．検索語 $\boldsymbol{q}$ の検索結果から「$\boldsymbol{q}$ や **s**」,「$\boldsymbol{q}$ の **t**」というフレーズを用いて検索語候補を取得する．検索語と取得した語をノードとしたグラフを提示し，ノードを選択したときにそのノードの語を検索語として再検索し，グラフを広げていくことで検索過程の“ 見える化 ”を行っている．本研究でも検索過程を利用するが，検索質問生成をユーザに任せるのではなく，自動生成し検索を行う点が異なる．

検索結果中に出現する特徴語の出現頻度の変化を用いて検索語候補を取得する研究[8]もある．そこでは，特徴語の出現頻度の変化を「増加」,「減少」,「高い値での推移」の 3 種類に分類し，出現頻度の変化がその特徴語への興味の変化と考え，興味の増加した語，減少した語，興味を維持している語が検索結果から抽出されている．さらに，興味の増減した語および興味が維持されている語の可視化手法を提案している．本研究でも同様に出現頻度の変化を興味の変化とみなして特徴語を抽出する．しかし，本研究は，文書クラスタリングを利用することで，絞込み検索においてユーザがまさに絞り込んもうとしている話題を抽出する点が異なる．さらに，検索語の想起支援ではなく，先読み検索を自動で行う点も異なる．

検索質問の変化を利用する研究には，検索質問の変化とクリックログからのフィードバックによって検索結果の再ランキングを行うものもある[9]．本研究は，検索結果の再ランキングではなく先読み検索を行う点で異なる．

2.3 文書クラスタリングによる話題の抽出

階層化クラスタリングを用いて，N 日分の電子番組ガイド（EPG）から話題の推移を抽出する手法[14]が提案されている．彼らはまず，EPG 文書を形態素解析し，各形態素の IDF 値を要素とする単語ベクトルを生成し，コサイン類似度を用いた階層化クラスタリングによって話題クラスタを生成している．また，各話題クラスタ内を文書の生起時刻を考慮して再クラスタリングすることで，話題内でのイベントの推移を抽出している．

森ら[15]はニュース記事を時間経過と共に話題の分岐や収束が起こるものとし，時間経過を考慮したクラスタリング手法を提案した．記事群を一定期間ごとに区切り，単語ベクトル間の類似度の計算に忘却定数を加味することで，クラスタリングを行っている．新たな記事と既存のクラスタ内の記事との類似度を計算したとき，新たな記事が複数のクラスタへ所属可能と判断されるとこれらのクラスタを併合することで，話題の収束を扱っている．また，期間の終了時に k-means 法によりクラスタの二分割を試み，条件を満たす場合に分割することで話題の分岐を扱っている．

我々も文書クラスタリングによって話題を抽出するが，話題を提示するのではなく，先読み検索に利用する点が異なる．

3. 提 案 手 法

本章で提案手法について説明する．本研究では，ユーザが目的とする検索結果を一度の検索で得られず，連続して検索を行う場面を想定して支援を行う．また本稿では，絞り込み検索を対象として検索支援を行う．ここで絞込み検索とは，検索質問に新たにキーワードを加えることにより，検索結果を絞り込む検索のことをさす．また，ユーザによる検索質問の入力を検索行動と呼び，一連の検索行動が n 回からなるとき，i 回目の検索の際に入力された検索質問を「i 回目のクエリ」と表現する．また，ユーザからの検索質問入力による検索を「通常検索」，ユーザの検索意図を推測して生成した検索質問によるシステムの自動検索を「先読み検索」と呼ぶ．以下に本稿で提案する先読み検索の流れを示す．

(1) 通常検索を行い検索結果を取得
(2) 検索結果から特徴語とその出現頻度を取得
(3) 前回の検索結果と今回の検索結果から特徴語の出現頻度の変化を計算
(4) 検索結果文書をクラスタリングし話題クラスタを生成
(5) 特徴語の出現頻度の変化を利用し検索意図に合う話題を特定
(6) 特定した話題を利用して先読み検索のための検索質問を生成
(7) 先読み検索を実行
(8) (1) に戻る

© 2010 Information Processing Society of Japan

Vol.2010-DBS-151 No.2
2010/11/12

検索結果の取得には Yahoo! ウェブ検索 Web API[⋆1]を用いる．

3.1 特徴語の抽出

特徴語の抽出は長畑ら[8] の手法を参考にした．取得した検索結果のタイトルおよびサマリから形態素解析器 Sen[⋆2]により形態素を取得する．得られた形態素から除外形態素リストに含まれる形態素を除き，残った形態素のうち，以下のいずれかの条件を満たす形態素を，特徴語を構成する要素として抽出する．

(1) カタカナのみから構成される形態素
(2) 名詞（形容詞語幹，副詞可能，非自立，代名詞は除く）
(3) アルファベットのみで構成される未知語
(4) 数字
(5) 漢字のみで構成される未知語
(6) 連体助詞の「の」
(7) 名前区切りの記号

上記の形態素が連続して現れた場合は連結する（**表 1**）．ただし，(6)，(7) は語頭にならない．連結処理により得られた語は長くなりがちであるため，接頭詞，接尾詞による分割処理を行う．具体的には接頭詞の直前と接尾詞の直後で分割する．たとえば，連結処理により得られた「東京都新宿区大久保」は**表 2** に示すように，形態素「都」,「区」が接尾詞であるため，その直後で分割され,「東京都」「新宿区」「大久保」となる．

以上の処理で得られた特徴語候補から，以下の語を取り除いた語を特徴語とする．

(1) 除外語リストに含まれる語
(2) ひらがなのみの語
(3) 数字のみの語
(4) アルファベット１字の語
(5) カタカナ１字の語
(6) 漢字１字の語
(7) 日時などを表す語
(8) 「ん」「ー」で始まる語

⋆1 http://developer.yahoo.co.jp/
⋆2 https://sen.dev.java.net/

表 1 連結の例
Table 1 Examples of concatenation.

形態素	連結後
道，の，駅	道の駅
ジョージ，・，ワシントン	ジョージ・ワシントン

表 2 分割の例
Table 2 An example of division.

形態素	品詞情報	分割後
東京	名詞–固有名詞–地域–一般	東京都
都	名詞–接尾–地域	
新宿	名詞–固有名詞–地域–一般	新宿区
区	名詞–接尾–地域	
大久保	名詞–固有名詞–地域–一般	大久保

3.2 文書クラスタリング

本研究では検索結果文書をクラスタリングすることで，ユーザが絞り込もうとしている話題を発見する．文書クラスタリングでは，まず検索で得られた文書集合中の文書をベクトルとして表現する．初期状態では各文書を全てクラスタとみなして，文書間の距離を計算する．距離の近い文書（クラスタ）を併合し，クラスタ間距離を再計算する．終了条件を満たすまでクラスタの併合を繰り返すことにより，最終的なクラスタを得る．

以下，3.2.1 項にて検索結果文書のベクトル表現について説明し，3.2.2 項で文書間距離を定義する．そして，3.2.3 項で本研究で採用したクラスタリング手法について説明する．

3.2.1 文書ベクトル

i 回目の検索によって得られた検索結果文書集合 D_i 中の j 番目の文書を以下のような文書ベクトルで表す．

$$\boldsymbol{d}_{ij} = \begin{pmatrix} w_{ij}(t_1) \\ w_{ij}(t_2) \\ \vdots \\ w_{ij}(t_n) \end{pmatrix} \tag{1}$$

ただし，$t_1, t_2, \cdots, t_n$ は i 回目の検索結果に現れる特徴語であり，$w_{ij}(t_k)$ はその重みである．特徴語 t_k の重み $w_{ij}(t_k)$ は，以下に示す TF-IDF による重み付けおよび，出現文書増加率による重み付けによって計算する．

© 2010 Information Processing Society of Japan

TF-IDF による重み

文書ベクトル $\boldsymbol{d}_{ij}$ 中の特徴語 t_k の重み $w_{ij}(t_k)$ を式 (2) で定義する．

$$w_{ij}(t_k) = tf_{ij}(t_k) \times idf_i(t_k) \tag{2}$$

$$idf_i(t_k) = \log \frac{n_i}{df_i(t_k)} \tag{3}$$

ここで，$tf_{ij}(t_k)$ は i 回目の検索の j 番目の文書における特徴語 t_k の出現頻度，n_i は i 回目の検索において取得した検索結果文書総数，$df_i(t_k)$ はそのうち特徴語 t_k の出現する文書数である．

出現文書増加率による重み

文書ベクトル $\boldsymbol{d}_{ij}$ 中の特徴語 t_k の重み $w_{ij}(t_k)$ を式 (4) で定義する．

$$w_{ij}(t_k) = \delta_{ij}(t_k) \times rdf\_increase_i(t_k) \tag{4}$$

$$\delta_{ij}(t_k) = \begin{cases} 1 & t_k\text{が文書 } \boldsymbol{d}_{ij}\text{に現れるとき} \\ 0 & otherwise \end{cases} \tag{5}$$

$$rdf\_increase_i(t_k) = rdf_i(t_k) - rdf_{i-1}(t_k) \tag{6}$$

$$rdf_i(t_k) = \frac{df_i(t_k)}{n_i} \tag{7}$$

ここで，式 (7) を i 回目の検索結果文書集合における特徴語 t_k の出現文書率，式 (6) を特徴語 t_k の出現文書増加率と呼ぶ．

3.2.2 文書間距離

i 回目の検索結果文書 $\boldsymbol{d}_{ia}$ と $\boldsymbol{d}_{ib}$ の非類似度として以下の 2 種類を採用し，3.2.3 項ではこの 2 種類の非類似度を距離とみなしてクラスタリングを行う．

ユークリッド 2 乗距離

$$dissim_e(\boldsymbol{d}_{ia}, \boldsymbol{d}_{ib}) = \sum_{k=1}^{n} \left(w_{ia}(t_k) - w_{ib}(t_k)\right)^2 \tag{8}$$

ベクトルのなす角

$$dissim_\theta(\boldsymbol{d}_{ia}, \boldsymbol{d}_{ib}) = \arccos\left(\frac{\boldsymbol{d}_{ia} \cdot \boldsymbol{d}_{ib}}{\|\boldsymbol{d}_{ia}\|\|\boldsymbol{d}_{ib}\|}\right) \tag{9}$$

ここで，$\|\boldsymbol{x}\|$ は $\boldsymbol{x}$ のノルム，$\boldsymbol{x} \cdot \boldsymbol{y}$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の内積を表す．

3.2.3 クラスタリング手法

文書クラスタリング法として以下の 2 種類の凝集型クラスタリングを採用した．また，生成するクラスタ数には一定の制約を課し，クラスタリングはこれを含む停止条件を満たしたときに停止する．

ウォード法

検索結果文書集合 D_i について以下の手順でクラスタリングを行う．

(1) 各文書 $\boldsymbol{d}_{i1}, \boldsymbol{d}_{i2}, \cdots, \boldsymbol{d}_{i|D_i|}$ を初期クラスタとする．

(2) 最小距離を与えるクラスタ A とクラスタ B を併合し，クラスタ C を生成する．

(3) クラスタ間の距離を再計算する．

- クラスタ C とクラスタ X の距離 R_{CX} は以下で計算される
- $R_{CX} = \alpha R_{XA} + \beta R_{XB} + \gamma R_{AB}$
  ただし，$|X|$ をクラスタ X の要素数として
  $$\alpha = \frac{|X|+|A|}{|X|+|C|}, \beta = \frac{|X|+|B|}{|X|+|C|}, \gamma = -\frac{|X|}{|X|+|C|}$$

(4) (2) へ戻る．

単純凝集法

検索結果文書集合 D_i について以下の手順でクラスタリングを行う．単純凝集法ではクラスタをベクトルで定義し，ベクトルの演算により，クラスタを併合する．

(1) 各文書 $\boldsymbol{d}_{i1}, \boldsymbol{d}_{i2}, \cdots, \boldsymbol{d}_{i|D_i|}$ を初期クラスタベクトルとする．

(2) 最小距離を与えるクラスタベクトル $\boldsymbol{A}$ とクラスタベクトル $\boldsymbol{B}$ を併合し，クラスタベクトル $\boldsymbol{C}$ を生成する．
併合されたクラスタベクトル $\boldsymbol{C}$ は以下の式で算出する．
$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$$

(3) 3.2.2 項の式によりクラスタ間距離を再計算する

(4) (2) へ戻る

クラスタリングの停止条件

クラスタの併合が l 回行われた時点で，以下の条件のいずれかを満たしたならば，そこでクラスタリングを停止する．

- $transition_{l+1}$ が $transition_l$ の N 倍以上になる
  ここで，$transition_l$ は l 回目の併合を与えるクラスタ間の最小距離と $l-1$ 回目の併合を与えるクラスタ間の最小距離との差分

© 2010 Information Processing Society of Japan

- $l+1$ 回目の併合によって，全文書の 50%以上を要素とするクラスタが生成される
- クラスタ数が全文書数（初期クラスタ数）の 5%以下になる

3.3 検索目的と合致する話題クラスタの選択

本節では，3.2 節の文書クラスタリングによって得られたクラスタの中からユーザの検索目的に合致するクラスタを自動選択する方法を説明する．本研究では出現文書率の増減をその特徴語への興味の増減とみなす．そこで，出現文書率の増加が大きい特徴語を多く含むクラスタをユーザの検索目的であるクラスタとみなし，逆に出現文書率の減少が大きい特徴語を多く含むクラスタはユーザの興味から外れたクラスタであるとみなす．そのため，各クラスタ内の文書中に含まれる特徴語について，式 (6) の値を合計したものをそのクラスタのスコアとする．すなわち，i 回目の検索において生成された，ある文書クラスタ C_m のスコアを式 (10) で定義し，この値が最も高いクラスタをユーザの検索目的を表す目的クラスタ $C_{purpose}$ とする．

$$score_i(C_m) = \sum_{\boldsymbol{d}_{ij} \in C_m} \sum_k \delta_{ij}(t_k) \times rdf\_increase_i(t_k) \tag{10}$$

3.4 先読み検索

3.3 節で得られた $C_{purpose}$ 内の文書の特徴語を用いて先読み検索の検索質問を生成する．まず，$C_{purpose}$ の文書に含まれる特徴語 t_k について，式 (6) を計算する．その値が大きい特徴語 s 個を，現在入力されている検索質問 Q_i に追加することで先読み検索用の検索質問 $Q_{predict}$ を生成する．以降，$Q_{predict}$ を先読み検索クエリ，$Q_{predict}$ に追加した特徴語を先読み検索語と呼ぶ．

$$Q_{predict} = Q_i \text{ AND } (t_1 \text{ OR } t_2 \text{ OR} \ldots \text{OR } t_s) \tag{11}$$

4. 実　　装

図 1 は実装したプロトタイプシステムの実行画面である．この例ではまず，ユーザは「GDP」で検索した後,「GDP 略語」で検索した．このとき，通常検索と同時に 3 章で説明した先読み検索クエリを用いて先読み検索を行っており，ユーザは「先読み検索」のタブを選択することで図 1 のように先読み検索の結果を閲覧できる．なお，画面左には先読み検索語の候補とそのスコアが表示され，上位の s 語が先読み検索語として使用されている．

また，実行時間は通常検索のみの場合 3 〜 4 秒であり，先読み検索は 4 秒前後である．先読み検索は通常検索のバックグラウンドで行われるため，先読み検索を含む 1 回の検索が完全に終了するには 6 〜 8 秒を要する．

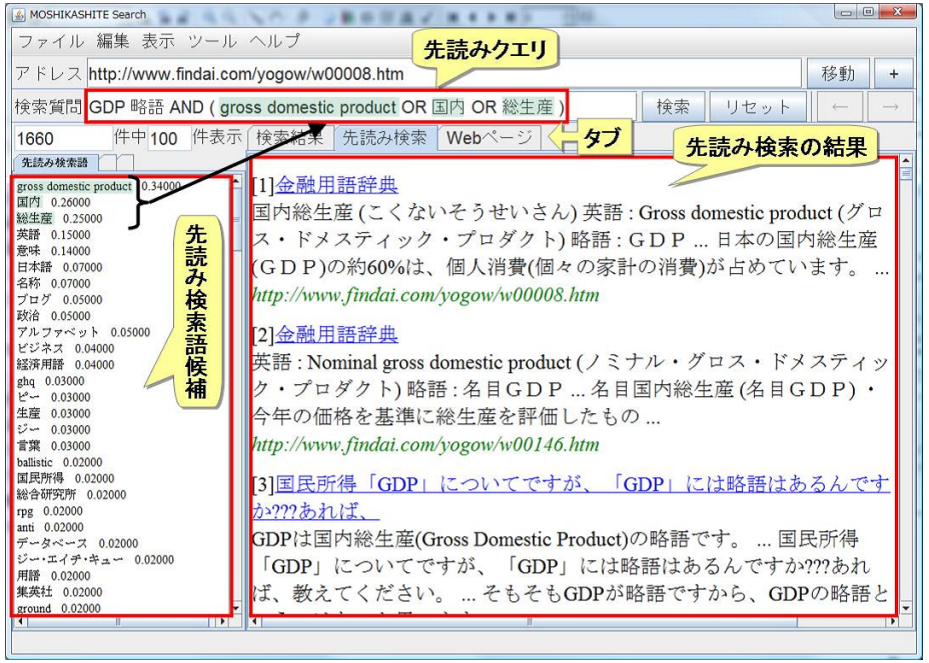

図 1 先読み検索の実行画面
Fig. 1 Screenshot of the prediction search.

5. 実　　験

5.1 パラメータ N の決定

クラスタリングの停止条件におけるパラメータ N を決定するため，予備実験を行った．本研究のクラスタリングでは，各クラスタを併合する際にクラスタ間の最小距離を計算する．よって，この最小距離が大きく変化したらクラスタリングを停止する．

図 2 にユークリッド 2 乗距離を用いたクラスタリングにおける，クラスタ間の最小距離の推移の一例を示す．横軸はクラスタの併合回数，縦軸は最小距離の値である．図 2 から分かるように，クラスタの併合が進むにつれて，クラスタ間の最小距離は増加していく．併合回数が３０から３１回の間と３４から３５回の間を赤線で示したが，この 2 箇所で最小距離が大きく増加している．よって，この例では併合が３０回行われた時点でクラスタリングを

© 2010 Information Processing Society of Japan

Vol.2010-DBS-151 No.2
2010/11/12

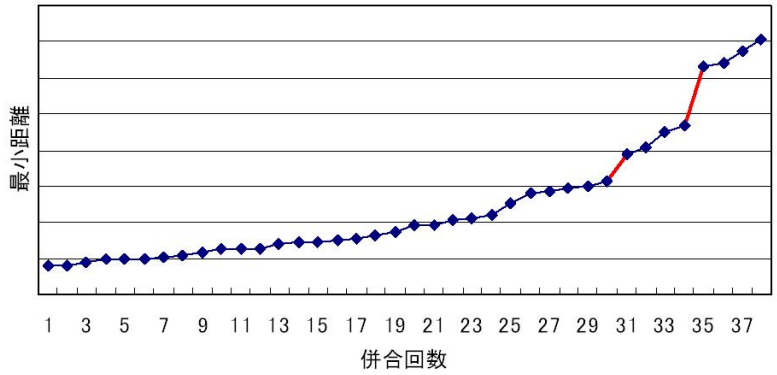

図 2 クラスタ間最小距離の推移
Fig. 2 Transition of minimum distances between clusters.

停止したい．本研究では，グラフの傾きが直前の傾きの N 倍以上になるとき，クラスタリングを停止するので，以降の実験によりパラメータ N を決定した．

実験は,「Yahoo!検索ランキング」の急上昇ワードランキング[⋆1]において，２０１０年７月５日～９日の期間のトップ１０の中から２語で構成される検索質問５例を選んで行った．これらの検索質問の２語目のキーワードを除いて検索した後，除いたキーワードを元に戻して検索することで擬似的に連続した検索を行い，3.2 節の方法でクラスタ数が１つになるまでクラスタリングを行った．その際の最小距離の推移を確認し，最初に傾きが５倍以上になる部分を求め，５例の平均を N の値とした．これを，**表 3** に示す手法１から手法８までの組み合せについて行った．決定した値を表 3 の右の列に示す．以降の実験では，この値を用いてクラスタリングを行う．

5.2 話題の抽出精度の評価

まず，3.3 節で抽出した目的クラスタ $C_{purpose}$ がユーザの検索目的をどの程度表しているかを評価する．実験に用いた検索質問および検索意図を**表 4** に示す．１回目のクエリ，２回目のクエリを用いて実際に検索を行い，手法１～手法８のそれぞれの手法で $C_{purpose}$ を得る．$C_{purpose}$ 内の文書について，著者が適合判定を行い目的クラスタ $C_{purpose}$ の適合率

⋆1 http://searchranking.yahoo.co.jp/burst_ranking/

表 3 提案手法の組合せとパラメータ N の値
Table 3 Combination of proposed method and values of the parameter N.

	文書ベクトルの重み	文書間の非類似度	クラスタリング手法	N の値
手法 1	TF-IDF	$dissim_e$	ウォード法	8.15
手法 2	TF-IDF	$dissim_e$	単純凝集法	9.27
手法 3	TF-IDF	$dissim_\theta$	ウォード法	8.13
手法 4	TF-IDF	$dissim_\theta$	単純凝集法	9.13
手法 5	出現文書増加率	$dissim_e$	ウォード法	7.87
手法 6	出現文書増加率	$dissim_e$	単純凝集法	5.25
手法 7	出現文書増加率	$dissim_\theta$	ウォード法	7.06
手法 8	出現文書増加率	$dissim_\theta$	単純凝集法	8.39

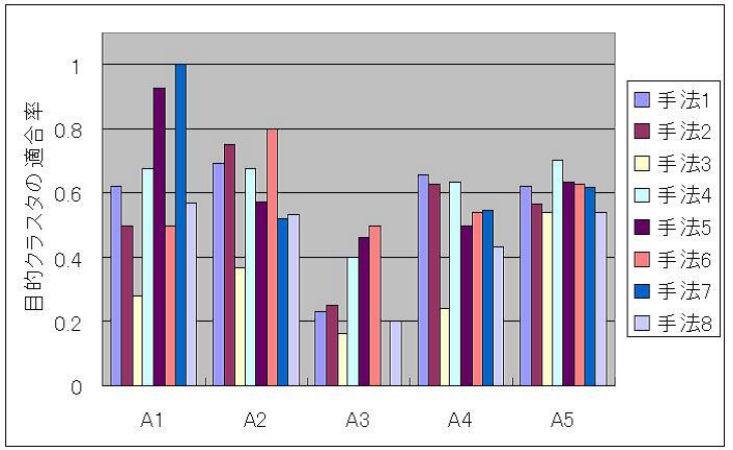

図 3 $C_{purpose}$ の適合率
Fig. 3 Precision of $C_{purpose}$.

を算出した．実験結果を**図 3** に示す．なお，実験は２０１０年９月１７日に行った．

図 3 より，A1 では高い適合率を示す手法が多く，逆に A3 では各手法が最も低い適合率を示している．つまり，A1 ではユーザの検索目的を表す文書をクラスタ $C_{purpose}$ として抽出できているが，A3 では上手く抽出できていないことが分かる．A3 では「Sleipnir のプラグイン」以外に「他のブラウザやプラグインと共に Sleipnir を紹介している文書」が $C_{purpose}$ に含まれていたため低い適合率となった．また，全体的に見ても，適合率が 0.4 以上 0.7 未満となることが多いことが分かる．

© 2010 Information Processing Society of Japan

表 4 検索質問と検索意図
Table 4 Queries and their search intention.

	1回目のクエリ	2回目のクエリ	検索意図
A1	GDP	GDP 略語	GDP が何の略語であるか知りたい
A2	口蹄疫 感染	口蹄疫 感染 人	口蹄疫が人に感染するかどうか知りたい
A3	Sleipnir プラグイン	Sleipnir プラグイン おすすめ	ブラウザ Sleipnir の便利なプラグインを知りたい
A4	Java エディタ	Java エディタ 無料	無料の Java エディタが欲しい
A5	熱中症 対策	熱中症 対策 家庭	自分でできる熱中症対策について知りたい

5.3 先読み検索の改善精度の評価

次に，我々が以前提案した先読み検索[3] との比較を行う．3) の手法では，クラスタリングは行っておらず，検索結果全文書中の特徴語について，出現文書率の増減および TF-IDF 法を利用したスコアを計算する．そのスコアに基づき，絞込み検索において興味の増加した語および興味の減少した語を抽出する．また，興味の減少した語を NOT 検索に利用することで，ユーザの検索意図と無関係の話題を検索結果から除外することを狙っている．

本稿の提案手法は 3) の手法の改良であるので，それと比較する．表 4 の検索質問を用いて検索を行い，先読み検索を実行する．$Q_{predict}$ に追加する特徴語数は共に３語とし，検索結果上位１０件の適合率 P@10 を算出した．実験は２０１０年９月１７日に行い，適合判定は岡山大学工学部情報工学科４年生の男性５名が行った．５名にはそれぞれ「高適合（５）」から「不適合（１）」までの５段階評価を行ってもらい，５名の平均が３より高い文書を適合文書と判定した．１回目のクエリ，２回目のクエリでの通常検索，3) の手法および，表 3 の８つの手法の P@10 を**表 5** に示す．

A1 では 3) の手法，提案手法の全てにおいてに “gross domestic product” が追加されていたため適合率が 1.0 になった．A2 については 3) の手法の方が高い適合率を示している．3) の手法が “口蹄疫 感染 人 AND (病気 OR 広辞苑 OR 影響) -拡大 -宮崎 -確認” という先読み検索クエリを生成したのに対して，提案手法では “口蹄疫 感染 人 AND (病気 OR 影響 OR 動物の病気)” や “口蹄疫 感染 人 AND (影響 OR 市場 OR 動物)” などが生成された．提案手法でも通常検索と同等の適合率は示しているが「人の移動により口蹄疫ウイルスが広まり，感染拡大を引き起こす」というニュースが含まれ，このニュースが適合率を下げていた．一方，3) の手法では “拡大” が NOT 検索に利用されているため，このニュースが検索結果に含まれず，高い適合率となった．A3 では，3) の手法では “Sleipnir プラグイン おすすめ AND (ie OR firefox OR opera) -公開 -フェンリル -スクリプト” という IE や Firefox などの競合するブラウザ名を含む先読み検索クエリを生成していたため，Firefox のプラグインを紹介しているサイトやブラウザを比較しているサイトを検索していた．さらに，“公開” や “フェンリル (Sleipnir とプラグインを開発している会社)” が NOT 検索されたため，Sleipnir の公式ページやプラグインの公開ページが検索されず，低い適合率になっていた．提案手法では “Sleipnir プラグイン おすすめ AND (ie OR タブブラウザ OR blog)” や “Sleipnir プラグイン おすすめ AND (タブブラウザ OR blog OR 特徴)” といった先読み検索クエリを生成した．競合するブラウザ名が減ったため，3) の手法より適合率が改善した．しかし，IE を追加することもあり，また Sleipnir の特徴でもある “タブブラウザ” を追加することで，Sleipnir 以外のタブブラウザについて比較紹介されているページが検索されてしまったため，通常検索の適合率を改善するまでには至らなかった．A4 では，提案手法の適合率が 0.6 と最も高く，その際の先読み検索クエリは全て “Java エディタ 無料 AND (コンパイル OR 入手 OR windows)” であった．3) の手法が “Java エディタ 無料 AND (開発 OR windows OR コンパイル) -eclipse -作成 -編集” という先読み検索質問を生成していたため，「Java によるエディタの開発」が検索結果に多く含まれ，さらに “eclipse” という Java の IDE が NOT 検索されていたため低い適合率となっている．これに対して提案手法では “入手” が含まれていたためエディタの入手法が紹介されているページを検索できていた．A5 では 3) の手法が “熱中症 対策 家庭 AND (医学 OR 冷凍 OR 運動会) -水分補給 -職場 -商品” という先読み検索クエリを生成していた．一方，提案手法で最も高い適合率 0.8 を示した手法１の先読み検索クエリは “熱中症 対策 家庭 AND (医学 OR ご協力 OR 学校)” であった．“水分補給” や “商品” といったキーワードが NOT 検索されず，熱中症対策が書かれている「学校から配布されるプリント」が多く検索されたため，高い適合率を示した．

6. ま と め

本稿では，絞込み検索を対象とした，先読み検索の改善手法を提案した．特徴語の出現頻

© 2010 Information Processing Society of Japan

表 5 提案手法の P@10
Table 5 P@10 of the proposed prediction search.

	通常検索		3) の手法	提案手法							
	1 回目の検索	2 回目の検索		手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	手法 7	手法 8
A1	0.4	0.7	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
A2	0.5	0.8	**1.0**	0.7	0.7	0.7	0.7	0.8	0.7	0.8	0.7
A3	0.3	**0.4**	0.1	0.3	0.1	0.3	0.3	0.3	0.3	0.1	0.3
A4	0.4	0.4	0.3	**0.6**	**0.6**	0.4	**0.6**	0.4	**0.6**	0.3	**0.6**
A5	0.5	**0.8**	0.5	**0.8**	**0.8**	0.5	0.1	0.4	0.5	0.4	0.5

度の増減を利用したクラスタリングにより，検索結果文書集合からユーザの検索目的に合致する文書を抽出し，抽出した文書中の特徴語を用いることで先読み検索クエリを生成した．評価実験では，クラスタリングによって抽出された目的クラスタがユーザの検索目的をどの程度表しているかを評価した．さらに，3) の手法と提案手法で先読み検索を行い，その検索結果の適合率を比較した．

評価実験の結果，提案手法は多くの場合 3) の手法よりも適合率の高い先読み検索を行うことができた．このことから，クラスタリングを用いて検索目的の文書を抽出することは先読み検索に有効であることを確認した．しかし，目的クラスタは適合率 0.4〜0.7 の間になることが多く，今後は目的クラスタの精度向上を目指していく予定である．

参 考 文 献

1) 藤田遼治，太田　学：検索質問と検索結果の推移に基づく先読み検索の提案，Web とデータベースに関するフォーラム（WebDB Forum）2009 論文集， 1A-3 (2009).
2) Ohta, M. and Fujita, R.: A Prediction Search Based on Changes of Queries and Search Results, *Proc. 4th International Conference on Ubiquitous Information Management and Communication*, pp.176–182 (2010).
3) 藤田遼治，太田　学：検索質問と検索結果の変化を利用した先読み検索，情報処理学会論文誌：データベース， Vol.3, No.3, pp.78–87 (2010).
4) 酒井浩之，大竹清敬，増山　繁：絞り込語提示による一検索手法の提案，言語処理学会第 7 回年次大会，pp.185–188 (2001).
5) 大石哲也，倉元俊介，峯　恒憲，長谷川隆三，藤田　博，越村三幸：関連単語抽出アルゴリズムを用いた Web 検索クエリの生成，電子情報通信学会論文誌 D， Vol.J92-D, No.3, pp.281–292 (2009).
6) 飯崎智之，志築文太郎，三末和男，田中二郎：検索過程を確認しながら条件指定が行える検索インタフェースの提案，第 19 回人工知能学会全国大会（JSAI 2005），3D1-02 (2005).
7) 中岡美華，大島裕明，田中克己：同位語発見システムを用いた子どもの利用履歴からの思考推移の抽出および分析，電子情報通信学会第 19 回データ工学ワークショップ（DEWS2008）， B7-1 (2008).
8) 長畑洋臣，太田　学：検索結果の推移の可視化による検索支援，Web とデータベースに関するフォーラム（WebDB Forum）2008 論文集， 5A-3 (2008).
9) Shen, X., Tan, B. and Zhai, C.: Context-sensitive information retrieval using implicit feedback, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.43–50 (2005).
10) 柘植　覚，獅子掘正幹，北　研二：サポートベクターマシンによる適合性フィードバックを用いた情報検索，情報処理学会研究報告， Vol.2000-NL-141, pp.83–88 (2001).
11) 山本岳洋，中村聡史，田中克己：ContextRank：語ベースフィードバックおよびそのコンテキストに基づく検索結果の再ランキング手法，第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM）2009， C8-1 (2009).
12) Rerank.jp. http://rerank.jp/.
13) Rerank. http://search.yahoo-labs.jp/rerank/.
14) 菊池匡晃，岡本昌之，山崎智弘：階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出，第 19 回データ工学ワークショップ（DEWS2008）， B3-3 (2008).
15) 森　幹彦：ニュース記事における時間変化する話題の抽出，第 22 回人工知能学会全国大会（JSAI 2008）， 2F2-02 (2008).

© 2010 Information Processing Society of Japan