D-024

ユーザの理解度に基づく検索結果の動的な個人化クラスタリング

Dynamic Personalized Clustering for Web Search Results Based on Users' Familiarity with the Topics

青野 壮志† 太田 学†
Hiroshi Aono Manabu Ohta

1．はじめに

Web 検索エンジンは広く利用されているが，必ずしも個々のユーザの嗜好に合った検索結果を返すものではない．その問題を解決する手段として，ユーザ個人の情報を利用した個人化検索と検索結果を分類・表示するクラスタリングの研究が広く行われている．

本研究では，ユーザの検索行動とブックマークを情報源として取得される嗜好情報を基に，個々のユーザに合わせて検索結果をクラスタリングするシステムを提案する．また，提案手法では，興味の遷移に合わせて変化する動的なクラスタを併せて提示する．

クラスタ選択行動からユーザの嗜好を読み取り，動的に検索結果リストを個人化する研究[3]があるが，本研究では，このような選択行動毎にクラスタリングの個人化を行う点が異なる．従来のクラスタリングシステムにおいて，ユーザの検索行動は，検索語に対応する静的なクラスタリング結果に依存していたが，動的に変化するクラスタを併せて提示することでユーザの検索行動の幅を広げている．また，本研究ではユーザの理解度（調べている話題についてどの程度詳しいか）に応じて，クラスタの抽象度を変化させる点に特徴がある．この理解度を考慮することで，ユーザにとって既知の情報や細かすぎる情報を排除できる．

2．提案する個人化クラスタリング

2.1 概要

提案システムは，入力された検索語に対する検索結果を Google から取得し，その検索結果から特徴語抽出を行う．抽出された特徴語はクラスタのラベルとなる．検索結果に抽出した特徴語が含まれていた場合は，その特徴語をラベルとするクラスタへ検索結果を割り当てるという非排他的なクラスタリングを採用している[1][2]．このとき，ユーザの行動情報とブックマークから作成されたプロファイルを用いて，クラスタリングを個人化する．また，短期的な嗜好情報と長期的な嗜好情報の両方を考慮することで，柔軟に個人化クラスタを別途生成する．システムの GUI 画面を図 1 に示す．図中①には検索語入力フォーム，②には初期クラスタリング結果，③には動的クラスタリング結果，④には選択したクラスタ内のページ集合が表示されている．

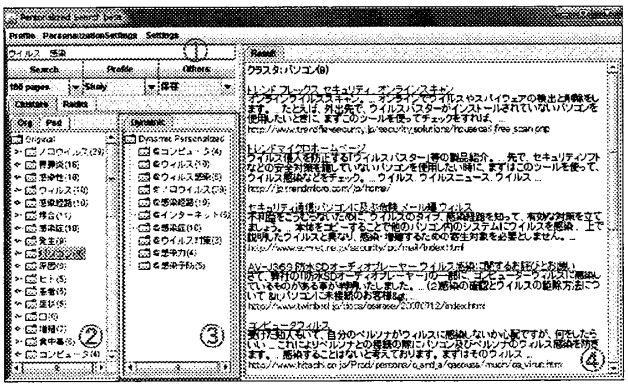

図 1：システムの GUI 画面

2.2 Web 検索結果からの特徴語抽出

本研究では，検索結果ページのドメイン名と，そのページのタイトルおよびスニペットに含まれる特徴語を，嗜好情報の基本単位として抽出した．ここで特徴語とは，クラスタのラベルとして利用できる語である．Sen[4]を用いて，IPA 品詞体系に基づき，次のように抽出ルールを設定した．【】内は，Sen で出力される品詞体系である．

(A) 特徴語候補となる語

- 【名詞】のうち【名詞－代名詞】および【名詞－非自立】を除いたもの
- 【未知語】のうちカタカナであるもの
- 【未知語】に含まれる英数字

(B) 文字数による制限

漢字ではない 1 文字の語は，それ単体では特徴をあまり表していないと考え，抽出対象から除外する．同じ理由から，英字のみの特徴語においては，先頭の文字が大文字の場合は 2 文字以上，小文字の場合は 3 文字以上のものだけを抽出対象とする．

(C) 語の前後関係

特徴語候補が連続する場合，それらを連結させてできた複合語を特徴語として抽出する．ここで，英文をそのまま特徴語として抽出することを避けるため，連続した英単語については，先頭と最後尾の語のみを抽出する．また，数字に関しては，前後に他の特徴語候補が存在した場合にのみ，その語と連結して抽出する．

(D) NG ワード設定

(A)～(C)の抽出ルールを適用しても，不適当な語が抽出されることがある．本研究では，「html」や「com」などの語を NG ワードとして設定し，これらの語を抽出対象から除外した．

2.3 プロファイルの作成

個人化クラスタリングのために，ユーザのプロファイルを作成する．プロファイルは，ユーザの行動情報とブックマークを情報源として，抽出された特徴語およびドメイン名とその重要度によって構成されている．本研究では，プロファイル内の重要度の高い特徴語とドメイン名がユーザの嗜好を表していると考える．作成されるプ

† 岡山大学大学院 自然科学研究科

表 1：作成されるプロファイルの例

	情報源			
	行動情報	重要度	ブックマーク	重要度
特徴語	音楽配信	100.0	情報処理学会	100.0
	着うた	73.4	パソコン	30.2
	ランキング	45.1	レシピ	24.4
	洋楽	35.5	ニュース	13.5
	：	：	：	：
ドメイン名	www.gyao.jp	100.0	www.ipsj.or.jp	100.0
	music.goo.co.jp	80.3	cookpad.com	64.6
	ja.wikipedia.org	76.0	kakaku.com	54.1
	m.oricon.jp	59.2	news.goo.ne.jp	23.5
	：	：	：	：

ロファイルの例を表1に示す．また，プロファイルの更新のタイミングと，その対象は次の通りである．

- **行動情報**
  - **(A) クエリ入力時**
    … クエリから抽出した特徴語
  - **(B) クラスタ選択時**
    … クラスタラベルに利用されている特徴語
  - **(C) Web ページ選択時**
    … タイトルとスニペットから抽出した特徴語
    … ページのドメイン名
- **ブックマーク情報**
  - **(A) ブックマーク登録時**
    … ブックマークしたページの本文全体から抽出した特徴語
    … ページのドメイン名

2.4 初期クラスタリング

提案するクラスタリング手法では，検索結果から抽出した特徴語がクラスタラベルとなる．クラスタの要素は，ラベルとなっている特徴語を含むページ集合である．

初期クラスタリングとは，検索実行後に最初に行われるクラスタリングのことである．本研究では，クラスタリング結果に対するユーザの選択行動に伴い，個人化クラスタを動的に生成する．初期クラスタの生成順を決定するために，特徴語 t の優先度 $priority(t)$ を検索結果の特徴 $R(t)$ と特徴語 t のプロファイルにおける重要度 $I(t)$ から算出する．

$$priority(t) = R(t) + I(t)$$

$$R(t) = \sum_{i=1}^{pnum}((2 * pnum - i) * df_i(t))$$

ここで，$df_i(t)$ は Google の i 番目の検索結果における特徴語 t の出現頻度，$pnum$ は検索結果ページ数である．$R(t)$ と $I(t)$ は共に最大値が 1 になるように正規化している．

2.5 動的クラスタリング

検索開始時点から現在までの行動情報と，第 2.3 節で示したプロファイル情報により，特徴語の優先度を決定し，個人化された動的なクラスタを生成する．以後，前者を「短期的な嗜好情報」，後者を「長期的な嗜好情報」と呼ぶ．この個人化クラスタは，ユーザがクラスタ選択またはページ選択といった行動毎に変化する．クラスタ選択の場合，これらの嗜好情報はクラスタ内のページ集合から取得する．動的クラスタの生成の流れを図 2 に示す．

提案する動的クラスタリングには以下の特徴がある．

- **嗜好に近いクラスタやページを選択した場合**
  ユーザは検索している話題についてある程度詳しいと考え，細分化したクラスタを生成することで，話題の詳細を与える．ユーザにとって，既知の情報を示すクラスタは生成しない．
- **嗜好と異なるクラスタやページを選択した 場合**
  ユーザは検索している話題についてあまり詳しくないと考え，抽象度の高いクラスタを生成することで，話題の概要を与える．ユーザにとって，細かすぎる情報を示すクラスタは生成しない．

2.5.1 2 つの嗜好情報

長期的な嗜好情報は，選択対象に含まれている特徴語と，それに対応するプロファイルの重要度により構成されている．この嗜好情報は，ユーザのこれまでの行動の蓄積であり，ユーザの長期的な嗜好を表している．

一方，動的クラスタの生成には，検索開始時点から現在までのユーザの興味の推移を示した短期的な嗜好情報も利用する．これは，選択対象に含まれている特徴語を情報源として作成される．プロファイル情報と同様に，行動毎に短期的な嗜好情報は更新される．ここでは，興味の鮮度を保つために，ユーザの嗜好の変化を考慮した更新を行う．ユーザの嗜好の変化の度合い C は，短期的な嗜好情報 S を用いて，以下のように算出する．

$$C = \frac{S\text{における選択対象の特徴語の重要度の合計}}{S\text{における全ての特徴語の重要度の合計}}$$

C が 1 に近づけば近づくほど，嗜好の変化は小さいと考え，S における特徴語 t の重要度 $S(t)$ を以下のように更新する．

$$S(t) \leftarrow S(t) * C + R(t)$$

ここで，$S(t)$ の最大値は 1 になるように正規化する．

2.5.2 動的クラスタ生成ルール

短期的な嗜好情報と長期的な嗜好情報に基づき，動的クラスタのラベルとなる特徴語の優先度を決定し，優先度の上位 50 個の特徴語を動的クラスタ候補とする．具体的なクラスタリングは次のように行う．

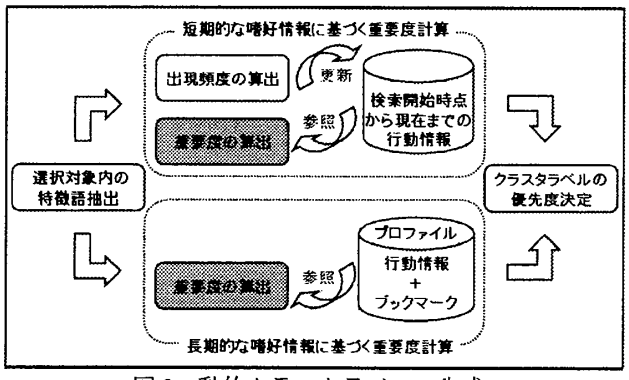

図 2：動的クラスタラベルの生成

(A) 選択対象に対するユーザの理解度

長期的な嗜好情報における全ての特徴語の重要度の合計を用いて，選択対象の 1 ページあたりの重要度を求める．その重要度をユーザの理解度とした．この理解度が低い場合，ユーザは選択したクラスタやページの話題についてあまり詳しくないとし，検索結果における出現頻度の高い特徴語だけをクラスタ候補とする．理解度が高い場合は，出現頻度によらず，大部分の特徴語をクラスタ候補とする．本研究では，出現頻度の高い特徴語を話題の概要を示す抽象度の高い語としている．

(B) 粒度の大きなクラスタの生成の抑制

ユーザの嗜好と異なるクラスタやページを選択した場合は，抽象度の高い粒度の大きなクラスタをそのまま生成すればよい．しかし逆の場合は，大小のクラスタ候補が混在しているが，この中から話題の詳細を示すような小さなクラスタのみを生成したい．そこで，粒度の大きなクラスタの生成を抑制するために，クラスタに含まれるページ集合間の包含関係を調べ，他のクラスタを包含するクラスタを生成対象から除外した．

3. 実験

3.1 実験方法

提案システムの有効性を確認するために 2 種類の実験を行った．実験 1 では，Discounted Cumulative Gain (DCG) [5] を利用して，初期クラスタリング結果がユーザの嗜好を反映しているかについて評価した．実験 2 では，動的クラスタの生成例を示してその有効性について考察した．

実験では，“ウイルス”と“感染”の AND 検索を実行した．この検索は，「パソコン」と「病気」の 2 通りの捉え方ができる．そこで，「パソコン」の関連語として，“コンピュータ”，“セキュリティ”，“プログラム”，“インターネット”，“メール”を，「病気」の関連語として，“症状”，“免疫力”，“食中毒”，“健康”，“予防”を設定し，それぞれの嗜好をもつ 2 人の仮想ユーザのプロファイルを次のように作成した．

(1) 関連語を検索語として入力する．
(2) 初期クラスタリング結果の上位 10 個のクラスタを選択する．
(3) Google 検索結果の上位 10 件のページを選択する．

3.2 実験 1：初期クラスタリングの個人化精度

クラスタリング結果の上位 20 個のクラスタを対象とした DCG を算出することで精度を評価した．算出式は以下の通りである．

$$dcg(i)=\begin{cases} g(1) & if(i=1) \\ dcg(i-1)+g(i)/\log_b(i) & otherwise \end{cases}$$

$$g(i)=\begin{cases} h & if(c(i)\in H) \\ a & if(c(i)\in A) \\ b & if(c(i)\in B) \end{cases}$$

ただし，ここで $c(i)$ は i 番目にランクされたクラスタを表し，$g(i)$ はクラスタ $c(i)$ の得点である．よって，$dcg(i)$ は上位 i 番目までのクラスタの累積得点を示している．クラスタにおける 3 段階の適合判定は，クラスタ内のページにより判定した．ユーザの嗜好と適合していると考えられるページが 8 割以上の場合を H，2 割未満を B，それ以外のクラスタは A とした．対数の底は $b=2$，配点は $(h,a,b)=(3,2,0)$ と設定し，ユーザの嗜好と適合しているクラスタが上位にランクするほど，DCG の値が高くなるようにした．

DCG による評価結果を表 2 に示す．さらに，各 DCG の値を，理想状態の DCG の値で割った結果（DCG の比）を図 3，図 4 に示す．本実験における理想状態は，上位 20 個のクラスタが全て適合度の高い，評価 H のクラスタになることであった．また，嗜好情報蓄積度とは，「パソコン」の嗜好をもつユーザの場合，関連語“コンピュータ”に対して，第 3.1 節の (1) ～ (3) により嗜好情報を取得したときを 1 とし，次の関連語“セキュリティ”による嗜好情報を追加したときを 2 としている．このように，5 段階のプロファイルを用いて実験を行った．0 は嗜好情報の蓄積がない状態である．

図 3 の「パソコン」のプロファイルでは，ユーザの嗜好情報が蓄積していくにつれて，ユーザが興味をもつと考えられるクラスタが上位にランクされるようになったことがわかる．一方，図 4 の「病気」のプロファイルでは，

表 2：初期クラスタリングの個人化精度

	パソコン関連の嗜好情報		病気関連の嗜好情報	
嗜好情報蓄積度	関連語	DCG	関連語	DCG
0	−	14.5328	−	16.7176
1	コンピュータ	17.6677	症状	18.0864
2	セキュリティ	19.4929	免疫力	17.6401
3	プログラム	19.5769	食中毒	18.6581
4	インターネット	20.3876	健康	18.1093
5	メール	21.0161	予防	17.5209

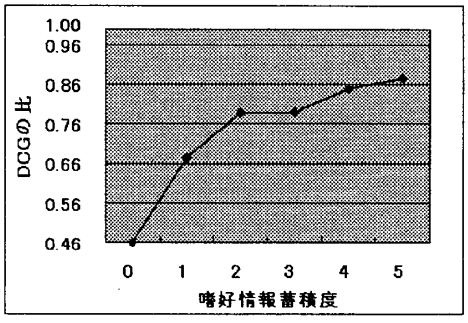

図 3：DCG の比（「パソコン」のプロファイル）

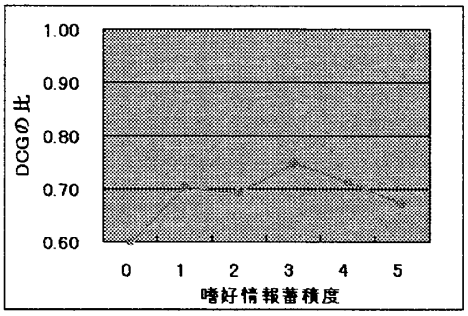

図 4：DCG の比（「病気」のプロファイル）

表 3：動的クラスタ（「パソコン」のプロファイル）

選択クラスタ	生成された個人化クラスタ
パソコン	インターネット，コンピュータウイルス， インストール，ネットワークウイルス， ノロウイルス，感染源，ウイルス対策， 原田ウイルス，感染済み，感染経路， PC ウイルス
トレンド マイクロ	パソコン，セキュリティ，Yahoo， ウイルス対策，使用，実行，注意，関連， コンピュータウイルス，詳細
インフル エンザ	原因，増加，パソコン，研究，予防， トレンドマイクロ，コンピュータ
感染症	原因，増加，パソコン， コンピュータウイルス，研究

表 4：動的クラスタ（「病気」のプロファイル）

選択クラスタ	生成された個人化クラスタ
パソコン	感染症，ノロウイルス，感染力
トレンド マイクロ	パソコン，ニュース，ウイルス対策，使用， 注意，症状，脅威，詳細
インフル エンザ	手洗い，原因，感染症，パソコン， ウイルス対策，PC ウイルス，子ども，肺炎， ノロウイルス
感染症	病気，インフルエンザ，手洗い，ピーク， 流行，最近，場合，予防，研究， RS ウイルス，PC ウイルス，肺炎， ノロウイルス

嗜好情報を用いない場合に比べれば精度は良くなったが，嗜好情報の増加がそのまま精度の向上には結びついていない．これは病気関連の嗜好情報の取得の際に，“検索”，“ページ”，“サイト”などの一般的な語が多く取得されたことが一因と考えられる．このような一般的な語については，NG ワードに設定するか，嗜好情報への影響を小さくすることが望ましいと考えられる．

3.3 実験 2：生成された動的クラスタ

作成した 2 種類のプロファイル（蓄積度 5）を用いて，どのような動的クラスタが生成されるかを調べた．“ウイルス”と“感染”で AND 検索したときの初期クラスタリング結果から，4 つのクラスタを順に選択したときに生成された動的クラスタを表 3，表 4 に示す．「パソコン」との適合度の高いクラスタから“パソコン”と“トレンドマイクロ”，「病気」との適合度の高いクラスタから“インフルエンザ”と“感染症”を表記の順に選択した．

表 3，表 4 の結果をみると，嗜好に近いクラスタを選択した場合，具体的なコンピュータウイルス名や病名がクラスタとして生成されており，話題の詳細を示す結果となったといえる．嗜好と異なるクラスタを選択した場合には，ユーザの嗜好に近いクラスタの中でも話題の概要を示すようなクラスタ（表 3 の“コンピュータ”や“コンピュータウイルス”など）が大分類として与えられている．また，「パソコン」および「病気」の両方に関連すると考えられるクラスタ（表 3 の“原因”や表 4 の“ニュース”など）も同時に生成されていることがわかる．これらは，多くの話題を含んだ抽象度の高いクラスタであり，ユーザに話題の概要を与えている．これらのことから，提案システムは意図したクラスタを生成していることが確認できた．

本実験では，嗜好に近いクラスタを選択した場合に多くのクラスタが生成される結果となった．これは，詳細なクラスタに加えて，抽象的なクラスタも同時に生成されたためである．これは，第 2.5.2 節の粒度の大きなクラスタの抑制が不十分であることと，両方の嗜好情報において重要度の高い“ウイルス”という単語（多義語）の影響により，嗜好との適合がうまく判断できなかったことが原因と考えられる．また，嗜好との適合がうまくいかなかったことが原因で，表 3 ではクラスタ“パソコン”を選択したときに，病気関連の話題である“ノロウイルス”が生成されたり，表 4 ではクラスタ“感染症”を選択したときに，パソコン関連の話題である“PC ウイルス”が生成されたりしている．

4．まとめ

本研究では，検索結果に特徴語およびドメイン名とその重要度によって構成されているユーザの嗜好情報を反映させる，動的な個人化クラスタリング手法を提案した．動的なクラスタの生成は，ユーザがクラスタ選択やページ選択を行う度に行った．実験により，この動的な個人化クラスタリングの有効性が確認できた．

本研究では，検索結果における出現頻度の高い特徴語を，話題の概要を示す抽象度の高い語であるとした．しかし，各特徴語の共起関係を調べ，様々な語と関連のある語を抽象度の高い語であるとする方法なども考えられる．よって今後は，話題の詳細・概要を示すクラスタについてさらに検討する予定である．

参考文献

[1] 川中翔：Web 検索結果の動的な個人化クラスタリング，岡山大学卒業論文 (2007).

[2] 成田宏和，太田学，片山薫，石川博：Web 文書検索のための非排他的クラスタリング手法の提案，DEWS2003 2-P-01 (2003).

[3] P. Ferragina and A. Gulli：“A personalized search engine based on web-snippet hierarchical clustering”，Proc. of WWW 2005, ACM Press, pp.801-810 (2005).

[4] 形態素解析器 Sen Project，http://ultimania.org/sen/

[5] K. Järvelin and J. Kekäläinen：“IR evaluation methods for retrieving highly relevant documents”，Proc. of ACM SIGIR 2000, pp.41-48 (2000).