5Z－4

概念体系を用いたFish Eyeマッチングによるユーザの視点の抽出*

高間 康史　石塚 満

東京大学工学部電子情報工学科

1 はじめに

インターネットに代表される，情報環境の急速な整備・拡大によって，研究や仕事などに必要となる情報を収集する過程はますます容易になりつつある．その反面，入手可能な情報量が人間の情報処理能力を越え，かえって効率が低下するという，いわゆる「情報過多（information overflow）」が問題となってきている．この様な事態を迎え，大量情報を扱う方法論の確立および，その過程を効果的に支援する知的システムの開発は急務であるといえよう．

大量のテキスト情報の取扱いに関しては，情報検索(IR)や発想法などの分野で研究されているが，これらの研究に共通する要素技術として，テキスト間の関連，特に類似性の発見があげられる．最も一般的なベクトル空間モデルにおいて，ユーザの興味をマッチングに反映したい場合には，適合フィードバックなどの手法を用いて各次元(単語)の重みの調整を行う．しかしこの方法では，ユーザの興味を正しく反映するためには十分な調整時間、訓練例が必要である事が指摘されている．

もっと大きな問題点として，重みの調整はユーザにとってブラックボックスであり，得られた重みベクトルからユーザの興味を読み取ったり，ユーザが修正を加えることは困難であることがあげられる．

我々が提案している**Fish Eye**マッチング[1, 2, 3]では，電子化辞書の持つ概念体系を利用して，特徴ベクトルの各次元を動的に生成、選択してマッチングを行う事により，ユーザのその時点での興味・視点を動的に，かつユーザに理解可能な形で反映する事が期待される．

本稿では，Fish Eyeマッチングの概要を紹介した後、ユーザがドキュメントを分類した結果から，ユーザの視点・興味を反映した特徴を生成する手法およびその具体例を紹介する．

2 Fish Eyeマッチング

従来のベクトル空間モデルでは，各単語を直交する成分として考えるため，類義語や反義語といった単語間の関係を利用することができず，重みの変更・調整のために訓練例が多量に必要となったり，あるいはユーザが明示的に計算・修正する事は非常に困難であるという欠点が存在する．

これに対し，我々の提案するFish Eyeマッチングでは，電子化辞書の概念体系を利用することによって，ユーザの視点に合わせ，特徴として使用する単語を動的に選択したり，あるいは複数の単語をまとめて一つの特徴として扱う事により，特徴ベクトル(**Fish Eye**ベクトル)を動的に生成する．

具体的には，全ての単語を直交成分とする基本特徴ベクトル$O_v(v_1,\ldots,v_m)$から，Fish Eyeベクトル$F_v(f_1,\ldots,f_n)$を計算する演算子として，下の二つの操作**Magnify, Shrink**を定義する．

$$Magnify(g_1,\ldots,g_n) \rightarrow \{f_i = v_j \ ; \ \exists k \ v_j \in g_k\} \quad (1)$$

$$Shrink(g_1,\ldots,g_n) \rightarrow \{f_i = \sum_j v_j \ ; \ v_j \in g_i\} \quad (2)$$

ここで，g_iはEDR電子化辞書の概念体系から求めた，単語の意味グループである．概念体系という，常識を利用する分，適合性フィードバックなどの純粋に統計的な手法と比べ，少ない事例からでも意味のある結果が得られる事が期待される．

また，得られた結果(意味グループ集合)からユーザの視点・興味を読み取ることが容易であるため，ユーザによる修正が可能であるだけでなく，発想支援システムへの利用にも適していると考えられる．

---

* Extraction of User's Viewpoint Using Fish-eye Matching Based on Concept Structure.

Yasufumi Takama, Mitsuru Ishizuka
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
takama@miv.t.u-tokyo.ac.jp

3 意味グループ抽出アルゴリズム

前節で定義した二つの演算子 Magnify, Shrink は，表裏の関係にあるといえる．すなわち，同じグループに対し，テキスト間の大まかな関係を概略視するには Shrink，ある視点・興味のみについて詳細に比較したい場合には Magnify を行えばよいと考えられる．従って，引数とすべき意味グループは，Shrink を元にした以下のアルゴリズムで求める事ができる．

ここで，ユーザの興味・視点に関連するテキスト集合を P，関連しないテキスト集合を N とする．また，基本特徴ベクトル空間を構成する全単語をリスト $Wlist$ に登録し，$i=0$ とする．

1. $Wlist$ 中の単語を一つでも含んでいるレベル i の意味グループを全て求め，その数をを m とする．$m=0$ の場合には終了．
2. 1. で求めたグループを次元とする Fish Eye ベクトルを N, P に属する全てのテキストから求め，次式に従って，各次元の重み $w_j(j=1\ldots m)$ を求める．値が負になったものについては 0 とした後，重みベクトル W を正規化する．

$$w_j = \frac{\alpha}{|P|}\sum_{d_k \in P} d_{kj} - \frac{1}{|N|}\sum_{d_k \in N} d_{kj}$$

3. グループ間の排他性に注意しながら，w_jの値がしきい値以上の意味グループについて，w_jの大きい順に抽出する．
4. 抽出された意味グループに属する単語を $Wlist$ から削除し，$i=i+1$ として 1. へ戻る．

ここで，下位に単語しか持たない意味グループをレベル 0 とし，下位にレベル $i-1$ 以下のグループしか持たないものをレベル i とする．また，グループ間の排他性ととして，同一の単語を含むグループは，二つ以上存在してはいけない，としている．

4 意味グループの抽出例

上記のアルゴリズムについて，インターネット上で配布されているニュース記事を題材として評価実験を行った結果の一例を紹介する．表 1 は，医学関係のニュース記事 7 つと，13 個のその他の記事について分類した結果に対し，上記のアルゴリズムを適用した事により抽出された意味グループの一部である．

この実験では，$\alpha=10.0$，しきい値は 0.01 としたところ，31 の意味グループが選択された．

表を見ると，blood, conges(tion), reduce, back, wind など，本来の意味と異なるグループとして選択されているものもあるが，比較的少数の記事からでも，ユーザの興味・視点を表すグループを抽出できていると言えよう．また，「記号」などの，有効でないグループも抽出されてしまうが，Fish Eye マッチングの特徴として，この様なグループの削除は人手で容易に行えると考えられる．

表 1: 医学関係の記事から抽出された意味グループの例

level	見出し	単語
1	記号	a, and, charac, d, f, l, n, s, t
2	性別で捉えた人間	blood, he, male, female, women
2	病気	bulimi, cancer, conges
3	薬品	chemic, drug, reduce, medica, laxati
3	臓器	back, cornea
3	動物の生理現象	infert, wind

5 最後に

本稿では，ユーザの興味・視点を反映した動的ベクトル生成／マッチング機構である，Fish Eye マッチングを紹介し，その意味グループ抽出アルゴリズムを提案した．

Fish Eye マッチングの利点として，ユーザの視点・興味を迅速に把握し，かつユーザにとって理解可能な形で提示する事があげられる．この様な特徴は，発想支援，特に収束的思考支援システムに代表される，情報整理支援システムに適すると考えており，実際にシステムを開発中である [3]．

今後の課題としては，システムの日本語化を含む，辞書の整備と，意味グループ抽出アルゴリズムの改良があげられる．特に意味グループの抽出に関しては，仮説推論などの AI 的手法が有効ではないかと考えている．

参考文献

[1] 高間, 石塚：情報整理・発想支援システムのための概念体系に基づく特徴ベクトルの動的生成・マッチング機構, 第 11 回人知全大論文集, pp. 372–373(1997).

[2] 高間, 石塚: 情報整理支援システムのための概念体系に基づく特徴ベクトルの動的生成, 第 55 回情処全大講演論文集, pp. 3-216–3-217(1997).

[3] 高間, 石塚：概念体系を用いた Fish Eye ベクトルの情報整理支援ツールへの応用, 人知研資 SIG-FAI-9702, pp. 97–102(1997).