

概念体系を用いた Fish Eye マッチングによる 視点を考慮した文書整理支援機能の実現

高間 康史 石塚 満

東京大学工学部電子情報工学科

〒 113-8656 東京都文京区本郷 7-3-1 東京大学工学部電子情報工学科石塚研究室
Tel. (03)3812-2111(ext. 6755)/E-mail. takama@miv.t.u-tokyo.ac.jp

あらまし インターネットに代表される情報環境の急速な整備・拡大により、研究や仕事などに必要となる情報の収集が容易になりつつある反面、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「情報過多 (information overflow)」が問題となって来ている。我々は、ホームページやオンラインニュースなどの文書情報の整理・熟読過程に焦点をあて、この過程を支援するツールの開発を行っている。本稿では、このための基盤技術として提案している Fish Eye マッチングについて報告する。Fish Eye マッチングはユーザの興味・視点を抽出し、外化できるようにベクトル空間モデルを拡張したものであり、この性質が文書整理支援機能に役立つものと考えている。

キーワード 特徴ベクトル, 概念体系, Fish Eye マッチング, 情報整理

Document Organization Function of Fish-Eye Matching with the Ability of Extracting User's Viewpoint Based on Concept Structure

Yasufumi Takama

Mitsuru Ishizuka

Dept. of Information & Communication Engineering,
School of Engineering, University of Tokyo.

Ishizuka Lab., Dept. of Information & Communication Engineering, School of Engineering,
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.

Tel. (03)3812-2111(ext. 6755)/E-mail. takama@miv.t.u-tokyo.ac.jp

Abstract

Recent rapid growth of information environment such as the Internet makes it easy for us to get vast information. On the other hand, "information overflow" is becoming a serious problem. In this paper, we propose a new mechanism for discovering a similarity among documents. This mechanism is called "Fish Eye Matching", which generates a feature vector from a text dynamically based on a concept structure derived from an electronic dictionary such as EDR. Using this concept structure, this mechanism can use semantic relations among words, which have not been considered in existing feature vector paradigms, and reorganize information space dynamically according to users' interests.

key words feature vector, concept structure, fish-eye matching, information organization

1 まえがき

インターネットに代表される情報環境の急速な整備・拡大により、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)」が問題となってきた。すなわち、今までは情報不足が知的活動の足かせとなっていたのが、今後は収集した情報をいかに生かすかが死活問題になるといえよう。

データベースからの知識発見はデータマイニング、数値データは統計的処理、といったように、大量データを扱う方法論についてはいろいろ研究されているが、我々にとって最も身近な情報源である文書情報が大量に入手された場合、これを活用する方法論については十分な研究がなされているとはいえないのが現状である。

我々は、インターネット等を通じて大量に収集された文書を熟読し、有効に活用するためには、各文書間の関係を整理しながら漸進的に読み進めていくことが有効であり、この過程を計算機によって効率的に支援するためには、その時点におけるユーザの興味・視点で捉えた場合の各文書間の関係を見出すことが必要であると考える。

この様な観点から、我々は **Fish Eye** マッチングと呼ぶ、概念体系を利用した新しい動的ベクトル生成・マッチング機構を提案している [高間 97]。Fish Eye ベクトルは、情報検索の分野で代表的なベクトル空間モデル [細野 94] の拡張モデルであり、単語間に存在する意味関係を利用して、ユーザの興味・視点を反映した特徴を動的に生成する。これによってベクトル空間モデルの持つ検索能力に加え、視点を外化することが可能となるため、ドキュメント間の類似性、差異を多様な視点から判断するのに有効であると考え。本稿では Fish Eye マッチングの概要について紹介するとともに、本手法の情報検索能力、視点外化の両能力を実験により示す。

続く2章では、従来の特徴ベクトルと比較しながら、提案する Fish Eye マッチングの概要について説明する。Fish Eye ベクトルを構成する意味グループの自動抽出法については3章で説明する。Fish Eye マッチングによる文献検索の評価実験について4章で示した後、最後に5章で他の研究との比較、検討を行う。

2 Fish Eye マッチング

2.1 文書整理支援

我々が論文やニュースなどを読んで考えをまとめたり、新たなアイデアを得ようとする場合、頭の中にある既存の知識と関連づけて整理していくと考えられる。この作業は、対象とする情報量が増えるにつれて困難な作業となってくるため、紙やコンピュータ上に各情報の関連を図示し、整理する事によって混乱を軽減する事が有効となる。

情報が形成する概念構造を外部に表現することは、既存の発想法 [星野 89, 川喜田 67] や発想支援システムの多くで採用されている。画面上にオブジェクトを配置・整理する過程を通じて概念の形成を支援するシステムは収束的過程支援システムと呼ばれており [國藤 93]、以下の様な効果が期待されている。

- 漠然と捉えていた自分の視点、疑問点を明確に意識することができる [遠藤 93, 篠原 93, Sugimoto 96]。
- 他者の構造との比較により、意見の相違や同意点、論点を明確にすることができる [Sumi 96, 女部田 96]。
- 図解された構造を変化させることにより、試行錯誤や発想の飛躍などが行いやすい [高杉 96, 杉山 97]。

これは、我々の目指す文書整理支援システム、すなわち「文書の整理を通じた熟読」をユーザに促すシステムにおいても重要な点である。あるテーマ、問題に関する文書を入手してみると、比較的短時間で思った以上に多量のドキュメントが集められる。しかし、これらを片端から読んでいくことは、考えただけで気の滅入る作業であり、精神的な苦痛を感じるだけでなく、一つ一つのドキュメントを雑に読んでしまうと考えられる。

これを避けるには、ドキュメント間の関係を意識しながら読み進めることが重要であり、今までに読み進めて来たドキュメント群が形成する概念構造から、現在自分が持っている興味・視点を把握しつつ、次に読むドキュメントを決定したり、構造に追加していくことにより、個々のドキュメントをより熟読できると考える。

しかし、ユーザの持つ視点・興味や、それに基づくドキュメント群の概念構造は最初からはっきり決まっているわけではなく、ドキュメントを読み進めていくうちに、変化しつつ、徐々に明確になってくるはずで

ある。その時点での興味の変化に伴い、今までに形成された概念構造も変化するし、次に読むべきドキュメントも異なってくるはずである。

ここで必要となるのは、視点により変化するドキュメント間の関係を捉える手法である。従来よく用いられて来たベクトル空間モデル (VSM) では、適合フィードバック [長尾 96] などを用いて、ユーザの興味・視点を反映したクエリベクトルを作成し、これと各ドキュメントから求めた特徴ベクトル (ドキュメントベクトル) との類似度 (内積) を計算する手法が一般的である。しかしこの方法ではユーザの興味がブラックボックスで扱われるため、自分の興味・視点が何であるかをユーザが明示的に知る事は難しい。

また、ベクトル空間モデルでは、各軸 (単語) 間の直交性が仮定されるが、実際には各単語間には類義語、反義語や上位・下位関係といった意味的な関係が存在するため、この仮定が満たされないという問題点が指摘されている。従って、ある視点から見て共通の特徴とみなせる単語群も、常に別々に扱われてしまうという問題がある。例えば、「車」という単語を含む記事と「自転車」という単語をそれぞれ特徴として持つドキュメントがある場合、乗物に関する記事という点では関連があるにも関わらず、ベクトル空間上では両ドキュメント間の関係を見出す事はできない。

以上の様な問題点を扱うため、本稿で提案する **Fish Eye マッチング** では、電子化辞書 の概念体系から抽出した意味の類似した単語グループ (意味グループ) を元に、視点に合わせた特徴を動的に生成してベクトルを構成し、マッチングを行う。詳細については以下に記す。

2.2 Fish Eye マッチングの定義

Fish Eye マッチングでは、ユーザの視点を反映したドキュメントベクトルを生成するために、通常の特徴ベクトルに対して単語の意味的關係を考慮した拡張を施して特徴ベクトルを動的に生成する。すなわち、ドキュメント中に含まれる単語をそのまま特徴として用いた基本特徴ベクトル (**Basic Feature Vector**) から、ユーザの視点において同様とみなせる単語は一つの特徴に縮退したり、その視点に関係のある単語のみを選択したりといった操作を加えることにより、ユーザの視点・興味の変化する毎に、それに応じた特徴ベクトル空間を動的に再構成する (図 1)。この、ユーザの視点を反映して動的に生成される特徴ベクトルのことを **Fish Eye ベクトル** と呼ぶ。

視点については、「そのドキュメントを読む際に仮定する、その属する話題、分野」であると考える。一つ

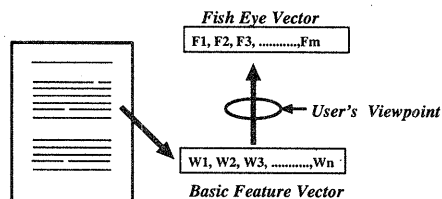


図 1: Fish Eye ベクトル生成のイメージ

のドキュメントの属する話題が一意に決まるということは稀であり、たいていは複数の視点から読む事が可能であると考えられる。また、ある種の話題間には包含関係が存在するため、同じ角度からの視点であっても、専門的な狭い話題として捉えたり、より一般的な、広い枠組で捉えたりする事も可能である。

話題としてどういったものかを考えるかであるが、本研究では EDR 電子化辞書の概念体系を利用する。すなわち、全ての話題には対応する概念が電子化辞書中に存在するものとし、その概念の範疇に含まれる単語を、特徴を構成する要素として採用する。

例として、図 2 について考えてみよう。ここでは 5 つのキーワードが存在しているが、概念体系に基づいて階層化すると、apple と lemon は fruits (果物) という概念の範疇に含まれ、これに tomato を加えた三つの単語は vegetable (野菜) という概念に含まれる。これらとは別に、bicycle と car は vehicle (車) という概念の範疇に含まれている¹。ここで、野菜に関する話題か、あるいは乗物に関する話題のどちらに属するかが区別できればいい程度の粗い視点であれば、apple, lemon, tomato を一つの特徴 vegetable に縮退し、car, bicycle も一つの特徴 vehicle に縮退して Fish Eye ベクトルを生成すれば良い。反対に、野菜の話題に集中し、野菜と果物に関する話題を区別して扱いたい場合には、apple, lemon を一つの特徴 fruits に縮退し、これと tomato を直交する特徴として用いれば良い。

話題に関係ある特徴のみをそのままとりだし、関係ない特徴を切り捨てる事は、概念体系のうちの興味ある部分だけをループで拡大してみることに相当する。従って、Fish Eye ベクトルは、基本特徴ベクトルに縮退 (**Shrink**) と拡大 (**Magnify**) という二つの操作を施して計算される。

単語集合を $W = \{w_1, w_2, \dots, w_n\}$ とすると、これから Magnify, Shrink の各操作によって得られ

¹ここで例示している概念体系は例のためのものであり、EDR 概念体系辞書から実際に抽出したのではない

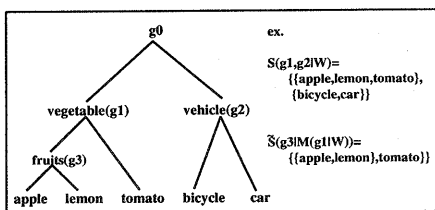


図 2: 概念グループの例

る Fish Eye ベクトルの特徴集合 $S(g_1, \dots, g_n|W)$, $M(g_1, \dots, g_n|W)$ はそれぞれ次式のように定義される。

$$S(g_1, \dots, g_n|W) = \{f_i | f_i = \{w_j | w_j \in g_i \wedge W\}, 1 \leq i \leq n\} \quad (1)$$

$$M(g_1, \dots, g_n|W) = \{w_i | w_i \in (g_1 \vee \dots \vee g_n) \wedge W, i = |(g_1 \vee \dots \vee g_n) \wedge W|\} \quad (2)$$

$$\tilde{S}(g_1, \dots, g_n|W) = S(g_1, \dots, g_n|W) \cup \overline{M}(g_1, \dots, g_n|W) \quad (3)$$

$$\overline{M}(g_1, \dots, g_n|W) = \{w_j | w_j \in W \wedge \overline{g_1 \vee \dots \vee g_n}\} \quad (4)$$

ここで、 g_i はある話題に関する単語のグループ (意味グループ) を表しており、特徴生成の際に背景知識の役割を果たすといえる。意味グループの計算方法については次章に記す。また、 $\tilde{S}(g_1, \dots, g_n|W)$ は、Shrink 操作の対象とならなかった W 中の単語についてはそのまま特徴として用いる操作であり、4章以降で利用するため便宜上定義した。前述の例に関する各操作について、図 2 に記してある。この時、 $W = \{apple, lemon, tomato, bicycle, car\}$ としている。

以上の操作により求められた特徴に基づいて、基本特徴ベクトル $O_v(v_1, \dots, v_n)$ から Fish Eye ベクトルを計算する際には、特徴 f に属する単語の、 O_v における値の総和を対応する値 v_f とする。ここで、 O_v の各要素の値については、TFIDF[長尾 96] などにより求められているものとする。

$$v_f = \sum_{w_i \in f} v_i \quad (5)$$

また、Fish Eye ベクトルに基づきマッチングを行う際には、通常の特徴ベクトルにおける場合と同様の指標を用いることができる。例えば次式の様に、ドキュメント d の話題に対する関連度 $Mag(d)$ はベクトルの長さ、ドキュメント d_i, d_j 間の類似度 $Sim(d_i, d_j)$ は両ベクトルの内積によって計算することができる。

$$Mag(d) = \sqrt{\sum_{i=1}^n f_i^2} \quad (6)$$

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^n f_{ik} \cdot f_{jk}}{Mag(d_i) \cdot Mag(d_j)} \quad (7)$$

3 意味グループの抽出

3.1 概念体系辞書からの意味グループの計算

本研究で使用している EDR 電子化辞書では、概念に関する辞書として、概念間の上下関係を規定する概念体系辞書と、概念間のそれ以外の関係を規定する概念記述辞書、および各概念を言葉で説明する概念見出し辞書が用意されている。このうち、本研究では概念体系辞書から意味グループを計算している。概念体系辞書によって整理される概念体系は、一般のシソーラスなどと同様に、各概念をノードとし、上下関係のあるものをノードで結んだ木構造の形に体系化される²。

前述したように、Fish Eye ベクトル生成演算子の引数として指定できる話題は、対応する概念が EDR 電子化辞書中に存在するとする。基本特徴ベクトルを構成する全単語について、概念体系を下層から上層へ辿りつつ、各概念ノードに対応する単語集合を求め、単語数が 2 から 256 の範囲に納まる概念ノードを意味グループとして用いている。この処理により、どの意味グループにも属さない単語が存在する可能性があるが、この場合には例外的に、単語数が 1 の意味グループとして扱う。

電子化辞書に記述された概念体系は、我々の常識を体系化したものと考えられるので、話題に対応した特徴を生成するために単語をグループ化する際の制約として適していると考えられる。

3.2 視点を表す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、同じ意味グループに対し、ドキュメン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 [高間 97]。従って、引数とすべき意味グループは、Shrink 操作を元にした以下のアルゴリズムで求めることができる。

²多重継承を許すため、厳密には木構造にならない

ここで、ユーザの興味・視点に関連するテキスト集合を D_P 、関連しないテキスト集合を D_N とする。これらは、今までに読んだドキュメント集合に対するユーザの分類といった形で得ることができる。また、初期化として、基本特徴ベクトル空間を構成する全単語をリスト $Wlist$ に登録する。以下で、 v_i は基本特徴ベクトル O_v 中の、単語 $word_i$ に対応する要素の値とする。

1. $Wlist$ 中の各単語 $word_i$ について、適合フィードバック [長尾 96] と同様の式により重み w_i を計算³。

$$w_i = \alpha \frac{1}{|D_P|} \sum_{O_v \in D_P} v_i - \frac{1}{|D_N|} \sum_{O_v \in D_N} v_i \quad (8)$$

2. $Wlist$ 中より、正かつ最大の重み w_k を持つ単語 $word_k$ を取り出す。なければ終了。
3. 意味グループ集合 $G_k = \{g_i \mid (word_k \in g_i) \wedge (\forall word_j \in g_i, word_j \in Wlist) \wedge (\forall word_j \in g_i, w_j \geq 0)\}$ を求める。これは、 $word_k$ を含むグループのうち、すでに抽出された他のグループと単語を一つも共有せず、かつグループ中に負の重みを持つ単語を含まないものを選択している。
4. (3) で求めた G_k 中の各グループ g_i について、重み W_{g_i} を次式にしたがって計算する。 $G_k = \emptyset$ の場合には (6) へ。

$$W_{g_i} = \frac{1}{|g_i|} \sum_{word_j \in g_i} w_j \quad (9)$$

5. G_k 中で、重みが最大のグループ g_i を抽出。
 $\{Wlist\} = \{Wlist\} - \{word_i \mid word_i \in g_i\}$ として (2) へ。
6. $\{Wlist\} = \{Wlist\} - \{word_k\}$ として (2) へ。

ここで式 (9) により、各グループに属する単語の重みの平均をグループの評価値 (重み) としている。これは、ベクトル空間モデルにおける「ユーザの興味に従った分類において、有効な指標となる重要な単語の重みは大きくなる」という仮定を拡張した、「ユーザの興味を表す概念には、重要な (重みの大きい) 単語が多く属している」との仮定に基づいている。すなわち平均値をとることにより、重要でない (重みの小さい) 単語が多く集まった意味グループより、少数でも重要な単語のみが集まった意味グループを優先して抽出している。

³ α は、正例、負例のバランスをとる適当な係数である

4 評価実験

4.1 実験設定

本稿で提案した Fish Eye マッチングについて評価実験を行った結果を記す。この実験の目的は、Fish Eye マッチングが、通常の特徴ベクトルと同程度の検索精度を維持しつつ、視点の外化を行える事の検証である。このため、インターネット上で公開されている、英語で記述されたオンラインニュース記事の中から、医学に関する記事を検索するタスクを行い、その適合率 (Precision) について調べるとともに、抽出された意味グループについて考察した。なお、ニュース記事が医学に関するかどうかの判定基準としては、その記事が公開されているニュースサイトにおける分類に従った。適合率 P は次式で与えられる⁴。

$$P = \frac{\text{検索文書中の該当文書数}}{\text{検索された文書数}} \quad (10)$$

検索方法であるが、正例および負例となるニュース記事をいくつか与え、3.2節に示したアルゴリズムにおいて抽出された意味グループ集合を引数として $\tilde{S}(g_1, \dots, g_n | W)$ 操作を行い、Fish Eye ベクトルを生成するとともに、抽出段階で得られた各グループに対する重み W_{g_i} (単語は式 (8) の値) を各要素の値としたものをクエリベクトルとし、ドキュメント検索を行った。

対象としたニュース記事の総数は 218 であり、そのうち該当文書である医学に関するニュース記事は 100 含まれている。これらの記事に含まれる単語からストップワードを除去し、EDR 英単語辞書に名詞の意味で登録されている単語を基本特徴ベクトルの構成要素として抽出したところ、1588 単語が抽出された。さらに、得られた単語集合から 3.1節で示したように意味グループを計算したところ、655 グループが得られた。

4.2 実験結果

テスト例として、正例のみ 5 から 15 記事まで変化させて与えた場合と、正例は一定数 (10) に固定し、負例を 0 から 10 記事まで変化させて与えた場合の適合率に関するグラフを図 3、図 4 にそれぞれ示す。グラフ中に示してある直線は、各条件における平均値を示している。また、横軸は検索された文書数、縦軸は適合率をとってある。また、比較のため、正例

⁴適合率と再現率の両指標を用いるのが一般的であるが、本章の目的は通常適合フィードバックとの検索精度の比較であり、検索文書数に大きく左右される再現率での比較は不要であると判断した。

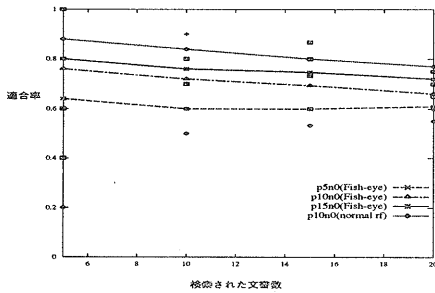


図 3: 実験結果 (1): 負例 0, 正例 5-15

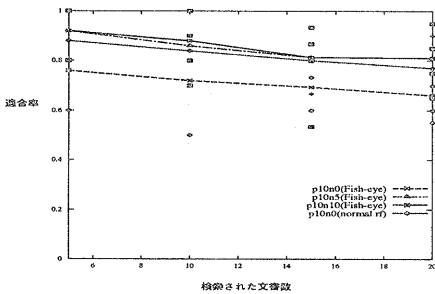


図 4: 実験結果 (2): 正例 10, 負例 0-10

10 記事, 負例なしの条件で通常の適合フィードバックを用いてクエリベクトルを作成し, 検索した場合を “p10n0(normal rf)” として両グラフに示している。また, システムの応答速度に関してであるが, Sun Ultra2 (メモリ 320MB) 上で実験を行った結果, 意味グループの抽出および文献検索までを含めても 5 秒以内であった。

図 3 より, 通常の適合フィードバックより適合率は若干落ちるものの, 正例を増やすにつれて適合率は上昇し, 適合フィードバックの性能に近づいていくことがわかる。この時, 正例の増加に連れて抽出される意味グループ数は増加していくのに対し, 意味グループ当たりの平均所属単語数は減少していくことが確認された。この事は, 正例の増加に伴い, 式 (8) によって計算される単語の重みがユーザの興味を正しく反映するようになるだけでなく, 抽出される意味グループが, 一般化しすぎたものから適度な粒度のものに分割され, 精度向上に貢献する事を表していると考えられる。これより, 概念体系から抽出された意味グループの, 特徴生成に関する制約としての妥当性, および 3.2 節で示したアルゴリズムの有効性が確認できる。

適合フィードバックより性能が劣る理由として主に

考えられるのは, 次の二点である。

1. 不適切な意味グループを特徴として抽出してしまっただけの誤差 (意味グループによるノイズ)
2. 意味グループ内に, 不適切な単語が含まれてしまっただけの誤差 (単語によるノイズ)。

(1) に関しては, EDR 電子化辞書から抽出された意味グループ中に適切なものが存在しなかった場合も考えられるが, それ以外の要因として, 単語を意味グループに縮退する際に, 一般化へのバイアスの方が強いことが考えられる。そこで負例を追加したところ, 図 4 に示した通り適合率が上昇し, 負例なしの場合の適合フィードバックより性能がよくなることが確認された。この時, 意味グループ当たりの平均所属単語数については, 正例を増やした場合と同様に, 負例の増加につれて減少したが, 今回は抽出される意味グループ数も減少することが確認された。これより, 負例の増加による適合率の上昇は, 意味グループの過度な一般化が抑制される効果に加え, 3.2 節のステップ (3) において不適切なグループが排除される効果も反映されたものと考えることができよう。後述するように, 文書整理支援システムの基盤技術として Fish Eye マッチングを用いた場合には, 負例が与えられるという前提は妥当であると考えられる。

残る (2) に関しては, 単語の持つ意味の曖昧さや品詞の違い, 単語それ自体の意味と, 複合語 (フレーズ) として使用された場合との意味の違いなどが原因であると考えられる。これについては, 複合語を特徴として考慮したり, 基本特徴ベクトルの計算時に構文解析などの技術を用い, 品詞の区別をして抽出するなどの対策が考えられる。この点に関連して, [猪股 98] では日本語ドキュメントへ対応するための改良が報告されているが, これは形態素解析を利用しているため上述の問題点がある程度解消されている事が期待できる。

また, 本実験において高頻度で抽出された意味グループの一部について, その見出し情報と, それに属する単語の一部を表 1 に示す⁵。これからも, 3.2 節で示したアルゴリズムの有効性が確認できる。また, 抽出された見出し情報や単語集合をユーザに提示することにより, 現在, どの様な視点・話題によってドキュメント間の関係を捉えようとしているかを知る手がかりをユーザに提供することができる。これは, 後述する文書整理支援システムへの応用を考えた際に, Fish Eye マッチングの持つ非常に大きなメリットであると言える。

⁵ ステマ後の単語の語幹を示している

表 1: 医学関係の記事から抽出された意味グループの例

概念 ID	説明	所属単語
3f98b3	健康状態の値	condit diseases sickne death health ...
444506	生体構成物質	protei immuno choles dna
30f6da	臓器	eye brain heart lung muscle knee ...
3f969e	病気	sympto syndro aids cancer cold ...
44479c	医薬品	drug medica medici laxati acid ...
30f6f7	医療器具	bandag cathet glasse contac patch

5 他の研究との比較

本稿で提案している Fish Eye マッチングについて、関連研究との比較検討を行う。

ユーザの視点を抽出し、可視化するという点では、多次元尺度構成法 [角 94] やバネモデル [高杉 96] などを用いて、画面配置上の位置関係の形で表現する方法がある。これらの方法は、本来多次元であるはずのドキュメント間の関係を、二次元空間での距離に変換する際に、その誤差が極小となる様に配置する。ここで、ユーザの初期配置によって最終的に達する局所解 (配置) が異なる、という性質を利用することにより、ユーザの視点を画面配置上に反映している。この方式は、ドキュメント数が増えると画面上が複雑になりわかりづらいうという欠点を持つものに対し、Fish Eye マッチングを用いた場合では、画面上での配置だけでなく、Fish Eye ベクトル生成時に使用された意味グループからも視点を読みとることが可能であるといった利点がある。

また、[Wulfekuler 96] で提案されている特徴クラスタリングでは、カテゴリが既知のドキュメント群に対し、各ドキュメントを次元とする空間において、K 平均分割クラスタリングによって特徴 (単語) をグループ化し、特徴空間を低次元化することが報告されている。このアプローチは Fish Eye マッチングの Shrink 操作と非常に類似しているが、純粹に統計的な操作であるため、ほとんど全ての単語を含むグループができてしまうなどの問題がある。この点でも、Fish Eye マッチングで用いる意味グループは人間にとって理解可能な概念に対応しているため、グループ計算時に概念体系がトップダウンな制約として有効に機能するだけでなく、得られた意味グループからユーザが意味を把握しやすいといったメリットがある。

6 むすび

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 マッチングについて提案し、実験によってその情報検索能力および視点の外化能力を示した。

単語を意味グループに縮退する事は、視点の外化

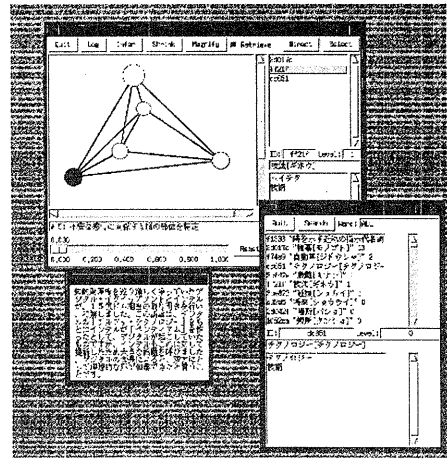


図 5: 情報整理支援ツール Fish View の概観

にしか意味がないわけではない。例えば、スポーツに関する記事に興味を持って文献検索を行う場合、単語を特徴として用いる通常のベクトル空間モデルでは、「サッカー」や「野球」といった、スポーツに関する単語の、クエリベクトルにおける重みが大きくなると考えられるが、検索された文書間の関係を見ようと思った場合、両単語は直交する特徴であるため、「サッカー」を含む文書と、「野球」を含む文書の間に、「スポーツ」という関連があることを捉えることができないと考えられる。これに対し Fish Eye マッチングでは、Magnify 操作によって話題を「スポーツ」に限定した上で、「サッカー」と「野球」を区別して扱うこともできるし、Shrink 操作によって両単語を一つの特徴に縮退することにより、「スポーツ」の観点からの文書間の類似性をとらえられる事が期待される。これについては後日実験し、報告する予定である。Fish Eye マッチングは文書整理支援ツールの基盤技術として提案したものであり、この様な効果が検証されれば、視点の外化能力と合わせ、より有力な支援が行えるものとする。

現在、Fish Eye マッチングを採用した文書整理支援ツールのプロトタイプを開発中である (図 (5))。現段階では、非常に限られた機能しか持たず、文書間の複雑な構造を扱うことができないが、今後は階層構造を許すなどの改良を加え、より扱いやすくする予定である。

参考文献

[遠藤 93] 遠藤 聡志, 大内 東: 統合型発想支援システ

- ム：FISM，人工知能学会誌，Vol. 8, No. 5, pp. 611-618(1993).
- [星野 89] 星野 匡：発想法入門，日本経済新聞社(1989).
- [細野 94] D. ELLIS (細野 公男監訳)：情報検索論：認知的アプローチへの展望，丸善(1994).
- [猪股 98] 猪股 健太郎，高間 康史，石塚 満：概念体系に基づく情報整理支援ツールの日本語化，第 56 回情処全大講演論文集，pp. 3-102-103(1998).
- [川喜田 67] 川喜田 二郎：発想法，中央公論社(1967).
- [國藤 93] 國藤 進：発想支援システムの研究開発動向とその課題，人工知能学会誌，Vol. 8, No. 5, pp. 552-559(1993).
- [長尾 96] 長尾 真：11. 情報の構造化と検索，自然言語処理，岩波書店(1996).
- [女部田 96] 女部田 武史，國藤 進：複数の KJ 法図解の差異や共通部を可視化する思考支援システムの実現と評価，人工知能学会研資 SIG-J-9602-6, pp. 28-33(1996).
- [篠原 93] 篠原 靖志：知識整理支援システム CONSIST-II，人工知能学会誌，Vol. 8, No. 5, pp. 593-600(1993).
- [杉山 97] 杉山 公造：発想支援ツール：ツール群の開発と統合化の試み - 創発メディア環境 - ，第 11 回人工知能学会全大チュートリアル(1997).
- [Sugimoto 96] Sugimoto, M., Hori, K. and Ohsuga, S. : A system to visualize different viewpoints for supporting researchers' creativity, Knowledge-Based Systems, Vol. 9, No. 6, pp. 369-376(1996).
- [角 94] 角 康之，堀 浩一，大須賀 節雄：テキストオブジェクトを空間配置することによる思考支援システム，人工知能学会誌，Vol. 9, No. 1, pp. 139-147(1994).
- [Sumi 96] Sumi, Y., Nishimoto, K. and Mase, K. : Facilitating Human Communications in Personalized Information Spaces, AAAI-96 Workshop on Internet-Based Information Systems, p. 123-129(1996).
- [高間 97] 高間 康史，石塚 満：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，人工知能学会研資 SIG-FAI-9702, pp. 97-102(1997).
- [高杉 96] 高杉 耕一，國藤 進：ばねモデルを用いたアイデア触発システムの構築について，人工知能学会研資 SIG-J-9602-7, pp. 34-39(1996).
- [Wulfekuler 96] Wulfekuhler, M., R. and Punch, W., F. : Finding Salient Features for Personal Web Page Categories, 6th Int'l WWW Conference, <http://www6.nttlabs.com/HyperNews/get/PAPER118.html>.