5L－1

自己組織化マップと適合フィードバック機能を用いた Web 文書群の分類ビュー機構について

波多野 賢治　佐野 綾一　田中 克己
神戸大学大学院自然科学研究科

1 はじめに

近年の World Wide Web (WWW) の発展に伴い，膨大な量の Web 文書の中から目的の文書を検索する，より柔軟な検索手法の要求が高まっている．従来の情報検索手法は，そのほとんどがユーザが検索目標を明確に持っている場合の手法であると考えられるが，WWW における情報検索では不十分であるため，本稿では，ユーザの多様な検索目的を支援するために，ユーザの検索エンジンに対する問い合わせ履歴，およびその問い合わせ時における適合フィードバック操作の履歴を利用した Web 文書群の分類ビュー機構を提案する．

2 適合フィードバック (Relevance Feedback)

検索システムを構築する上での大きな問題の 1 つは，ユーザがどのような検索意図を持っているときにどのような問い合わせを行うのか，あらかじめ予測できないため，ユーザが入力した問い合わせに対して求めているような検索結果が返ってくる保証がない．このような状況を解決する 1 つの手法に適合フィードバック (Relevance Feedback)[3] という手法があり，検索結果に対して適合，不適合を評価を反映させて，検索結果にフィードバックをかけながら検索を繰り返し，徐々に検索結果をユーザの求めるものに近づけていくというものである．

3 Web 文書の分類ビュー機構化

我々は，これまでに 自己組織化マップ (SOM)[2] を利用して情報の構造化や分類を自動化するシステムを開発してきたが [1]，ユーザ側の意図や興味をマップに反映することができないという問題があった．そこで，システムが生成したマップを適合フィードバック機能によりユーザとのインタラクションを行いマップを再構成することで，よりユーザに依存した検索を可能とした [4]．さらにユーザによる適合フィードバックの履歴を利用することで一種のユーザビューを構築し，これらの問題点を解消する工夫を行った．本システムの構成およびデータの処理手順を図1に示す．

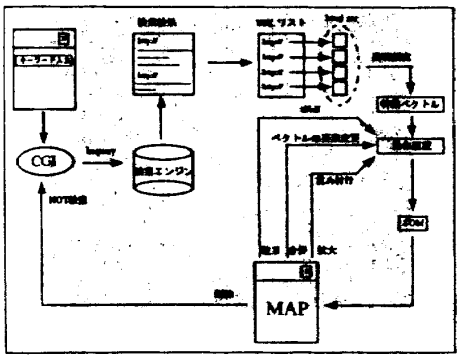

図 1: システムの全体構成

3.1 ユーザインタフェース

SOM は，入力データ群をトポロジカルマッピングの性質により 2 次元空間に表示することができるため，マップを生成するには大変有効な手段であると考えられる．そこで，我々は SOM に VRML ベースのインターフェースを採用し，マップ中のセルを選択することでそのセルに分類されたマッチページの URL とセルに含まれる 3 つのキーワードのデータが WWW ブラウザで参照できるように設計し，これに適合フィードバックの機能を付加することで，より検索結果の URL マップの分類状況が把握しやすいようになっている (図2参照)．

こうして得られた URL のマップに対して，ユーザの意図や興味を反映させる機能は，ユーザが欲している情報を効率的に検索する上で大変重要な機能である．本システムにおいては，Web 文書分類のための適合フィードバック機構を実現するために，以下の機能を組み込んでいる．

- 領域の拡大

  マップ上のあるキーワードに関係のあるマッチページをマップ上に浮かびあがらせる操作を行う．

- 領域の削除

  マップ上のあるキーワードに関係するマッチページは不要だと判断した場合に，その領域の削除を行う．検索エンジンに対して NOT 検索を行っている．

- 領域の合併

"Classification Views for Web Documents Based on Self-Organizing Maps and Their Relevance Feedback"
Kenji Hatano, Ryouichi Sano, and Katsumi Tanaka.
Graduate School of Science and Technology,
Kobe University.

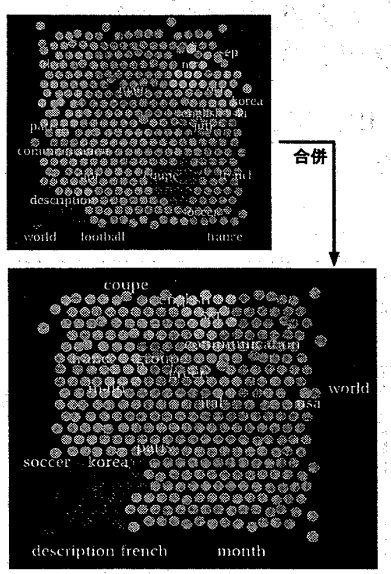

図 2: 適合フィードバックの一例

マップ上の複数のキーワードを 1 つに合併することで，ユーザの見やすいマップを再構成する．

- **領域の注目**

  マップ上の必要なキーワードを選択することにより，選ばれた領域内にあるマッチページのみで新たなマップを生成し，ユーザの欲するマッチページを絞り込んでいく．

- **領域の分解**

  マップ上の不明なキーワードを選択することにより，そのキーワードによる領域形成を行わず，他のキーワードによりマッチページの分類が行われる．

また，過去に行ったフィードバック操作の履歴を利用しすることで，同類の問い合わせに対して過去の適合フィードバック操作を利用した分類ビューをユーザに提供でき，ユーザ独自のビューを生成することが可能となっている．

3.2 評価

本研究の評価のために，検索精度の評価に情報検索で一般的に用いられている「精度 (適合率)」と「再現率」を測定する．本実験における「精度」と「再現率」を次のように定義する．

全てのページのの中から，マップ上において n 個の領域を作っているキーワード $k_i(i=1,\cdots,n)$ を特徴として表すページの数を $similar(k_i)$ とし，k_i の領域に分類されているページの数を $neighbour(k_i)$ で表わすとすると，精度は，

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\frac{|neighbour(k_i)\cap similar(k_i)|}{|neighbour(k_i)|}$$

で表わされ，また再現率は，

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\frac{|neighbour(k_i)\cap similar(k_i)|}{|similar(k_i)|}$$

で表わされる．

今回の様に検索エンジンの出力結果というものは，「情報が適合する，しない」という評価は主観的であり，ユーザによって異なるものであるので，ここでは第一著者の基準でページの内容の評価を行った．検索エンジンの出力結果に適合フィードバックを施す前と施した後の精度と再現率の変化を表 1 に示す．

表 1: 精度と再現率の評価

	精度 (%)	再現率 (%)
前	37.44	34.50
後	48.82	44.84

表1が示しているように，フィードバックを行うことにより精度および再現率が上昇しているため，検索時に目的の情報にたどり着くまでのユーザの労力が軽減できており，これらのフィードバック機能の有効性が確かめられている．しかし，領域ごとの精度および再現率の比較は行っているものの，適合フィードバックによる構成領域 (キーワードの個数) の変化について考慮していないため，新たな評価尺度を考える必要がある．

4 おわりに

今後は Web 文書の構造情報を考慮した複数の Web 文書を基本単位とした Web 文書群の分類手法についての考慮，および質問再構築を行う適合フィードバック機能を考える必要があると思われる．

謝辞 この研究は，一部，日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」および文部省科学研究費重点領域研究「高度データベース (No.275)」(課題番号 08244103) による．ここに記して謝意を表す．

参考文献

[1] K. Hatano, Q. Qian, and K. Tanaka. A SOM-Based Information organizer for text and video data. In *Proc. of the 5th International Confarence on Database Systems for Advanced Applications (DASFAA'97)*, pp. 205–214. World Scientific, Apr. 1997.

[2] T. Kohonen. The Self-Organizing Map. *Proceedings Of The IEEE*, Vol. 78, No. 9, pp. 1464–1480, 1990.

[3] G. Salton. *Automatic Text Processing: The Trancefor-mation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[4] 佐野綾一，波多野賢治，田中克己．自己組織化マップを用いた Web 文書の対話的分類とその視覚化．情報処理学会研究報告，98-DBS-116-5, pp. 33–40, Jul. 1998.