情報検索における文書集合の組織化について†

1Ｔ－7

○田中 栄治　宮崎 哲夫　古城 則道‡
学習情報通信システム研究所§

はじめに

近年のインターネットブームにより，誰もが世界中の様々な情報(リソース)にアクセスすることが可能となった．インターネットは，誰もが自由に情報を発信することができるため，様々な人々が多種多様な情報を発信している．しかし，それらの情報はテキストデータでさえ定型化されてはおらず，膨大な情報のほとんどが未整理のままであるといえる．

このような傾向は，計算機（ネットワーク）の発達前に比較して，個々の能力を大きく高める可能性を持っている反面，情報の氾濫によりユーザにとって本当に意味のある（必要な）情報にたどり着けない，いわゆる「情報の過負荷（*Infomation Overload*）」という問題を引き起こしつつある．この情報の過負荷という問題が，創造的思考のいわば予備段階である情報収集に過程において，大きな障害となるであろうことが指摘されている [4].

これらの大量のリソースを人手を介することなく自動的に分類・管理できれば（組織化できれば），有益な情報の入手や検索が容易になり,「効率的な情報活用」が可能となるのではないであろうか．

本稿では，できるだけ人手を介さずにユーザが求めている情報を提供できる自動分類システムの構築を行なう．このシステムは，動的なクラスタリングと，分類を依頼するユーザからの情報を用いることを特徴とする．

組織化の必要性

もし計算機が非常に賢くなってメッセージの内容を理解することができるようになれば，情報の受け手にとって意味のあるものだけを選び出したり，あるいは緊急に対応すべきものから順に提示するといったことができるようになるであろう．このように情報の洪水の中から本当に意味のある情報のみを透過させるようなシステム（フィルタ）があれば，情報の過負荷の問題は克服される．

また，計算機に内容理解を行わせることがまだ可能でない現在では,「情報の構造化」を行うことによって，ある程度情報の過負荷を軽減させることができる．つまり，情報を構造化し更にコミュニケーションの枠組を与えることにより，計算機はメッセージの内容そのものを理解しなくとも情報の組織化を行うことが可能となる．

しかし，現在のインターネットの性質を考えると，このアプローチは有効であるが，現実的な解決策ではないことがわかる．インターネットリソースはそのほとんどが構造化されてはおらず，そのような規定も存在しないのである．

情報の送り手側に組織化のための特別な情報を付加する必要が無く，また，情報の受け手側も特に組織化のためのルールやフォーマットなどを指定する必要が無いようなシステムを構築することにより，情報の氾濫が起きても受け手側にとって興味のある情報だけを抽出することが可能となる．

また，文書集合の組織化を，受け手が行うのではなく誰か（分類者）に組織化を依頼することによって実現できれば，受け手側に負担が生ずること無く，膨大な情報の中から必要な情報を透過的に入手することができる．

文書の分類方法についての考察

本研究の目的は，ユーザが膨大な情報の中から，必要な情報に透過的なアクセスを可能とするような環境を構築し，その情報を組織化する事により効率的な学習環境を提供する事にある．

そのため，計算機による文書の自動分類を検討し，分類する文書の知識を持たない人間が，どのようにして分類を行い，何をすれば，正しい分類結果に近づけるかについて考察する．

そこで，我々は人間による文書分類実験を行ない，文書カテゴリに対する知識の有無や教示による分類結果の変化について検討を行なった [2].

結果を考察すると，やはり知識の有無による分類結果の違いは大きく，知識を有する人間の方が，より正確に分類を行っていることが分かる．また，知識の無い者が文書分類の依頼を受けた場合，依頼者からの要望に応じるためには，その文書に書かれている言語種やカテゴリについての知識を与える“教師”が存在すれば，分類精度が向上する事も確認する事ができた．

実験から，どのような知識を保持していても，人間が頭の中で構文解析や意味解析を行うとは限らず，どちらかと言えば，共通の単語が多く含まれているなどのキーワード情報を基にした分類方法に頼る場合が多いということが分かる．これは，人間は認知的負荷の低い方法，すなわち，

---

†An Approach on Document Systematization for Information Retrieval
‡Eiji TANAKA, Tetsuo MIYAZAKI, Norimichi KOJO
§Software Research Laboratory

Table 1 人間の分類法とシステムの対比

	人間の分類	システムの機能
1	文書を取り出す	
2	単語の抽出	キーワード候補の作成
3	キーワードの抽出	
4	文書の抽象化	重要度の算出
5	文書の特徴抽出	文書のベクトル化
6	クラスタリング	

常にあまり思考する必要がない方法を選択しているといえる.

システムの構築

文書分類が，文書の意味を理解できなくとも可能であるならば，現在のコンピュータでも文書の分類はできるはずである．文書の自動分類は，電子化されたテキストデータを基に分類を行うが，もちろん文書の内容を理解する訳ではない．文書の中で使われている重要なキーワードを抽出し，抽象化，特徴付けを行なうことにより，分類を可能としている．

我々が定義した分類法を，計算機上の機能と照らし合わせると，**Table 1** のようになる．システムの検証には，人間における分類実験に用いたものと同様な文書を使用した．

また本研究では，電子化文書を分類するエンジンとして SCONN (*Self-Creating and Organizing Neural Networks*) を用いている [1]．SCONN はニューラルネットワークで用いられる手法を基にしており，神経細胞を模倣しながらも，より少ない PE 数で有効な認識ができるような教師なし学習型モデルである．

我々が，分類アルゴリズムに SCONN を選択した理由は，

- 複雑な処理を行わず，距離的に‘近い’ものをクラスタリングしていく
  → ‘意味的’な処理を使わず，‘視覚的’に近いパターンをクラスタリングできる，今回定義した人間の分類法に類似している．
- 逐次的に処理を行うため，インタラクティブなクラスタリングが可能である
  → クラスタ内にある他の文書に依存しないため，情報の追加 / 削除などに柔軟な対応が可能である．

という特徴を持っていた点である．

検討および考察

本手法を用いて行った分類結果を **Fig. 1** に示す．結果を考察すると，定性的に知識情報を持たない者が行うよりも良い分類を行なっていることが判明した．

これは，文字レベルではなく単語レベルでキーワードを抽出し，キーワードラベルとキーワードの出現場所による重要度の定義という知識を用いている事による．

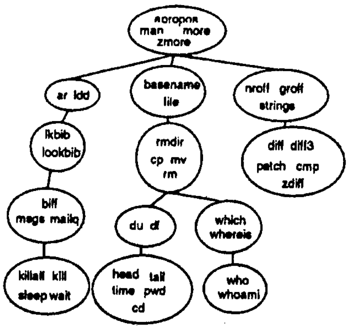

Fig. 1 本システムを用いた文書の自動分類

本研究では，計算機の発達に伴って生じた情報の過負荷問題を，計算機自身によって解決させるために，人間の手法を模倣し情報の整理・分類を自動化させるシステムの構築を提案した．

その結果，人間が行っている分類手法のモデル化の有効性を確認できた．まだ，機能として実現できていない部分もあるが，プロトタイプを作成し，それを用いた検証実験から，人間が行った分類結果との定性的な一致を認めることができた．

このことにより，人間の処理過程を抽象化しそれを計算機上でシミュレートすることの有用性を確認することができた．

今後は，人間の文書分類に対してのさらなる検討と，抽象化や特徴付けの処理の自動化を実装する予定である [3]．また，分類精度の向上についても随時検討していきたい．

参考文献

[1] Doo-Il Choi and Sang-Hui Park : Self-Creating and Organizing Neural Networks, *IEEE TRANSACTIONS ON NEURAL NETWORKS*, Vol. 5, No. 4, pp. 561–575 (1994).

[2] 田中 栄治, 宮崎 哲夫, 古城 則道：文書の組織化手法を用いた学習者モデル構築への一考案, *1996* 年電子情報処理学会ソサイエティ大会予稿, (1996).

[3] 宮崎 哲夫, 田中 栄治, 古城 則道： 文章の意味空間へのマッピング, 情報処理学会第 *53* 回全国大会予稿, (1996).

[4] 渡部 勇： 緩い協調：協調情報フィルタリングシステム，情報処理学会研究会報告 (ヒューマンインタフェース), Vol 35, No. 24, pp 179–186 (1991).