3D−7

仮想 Newsgroup の導入による NetNews の記事探索及び分類について

三谷 和史†，南 弘征‡，宮本 衛市†
北海道大学 工学部†，小樽商科大学 社会情報学科‡

1　はじめに

Internet などで広く普及している NetNews の流通量は膨らむ一方で，現在は約 150Mbyte/day と言われている．ユーザがこれら全てに目を通し，内容に応じて区分を行なうことは事実上不可能となってきた．

本研究は，不要な情報へのアクセスをできるだけ回避すること，必要な情報の漏れを少なくすることを目的として，NetNews の記事相互間について類似度を想定し，統計学的手法に基づいて作成される仮想的な距離空間に配置を行なったのち，自分の必要とする情報『周辺』の情報を収集することを目的としたものである．

2　情報の分類と選択

一般に，情報は関連性を持たずに存在している．そのため，利用しようとした際には所在を発見できなかったり，参照関係などがわからないなど必要以上の手間がかかる．そのため，我々は日頃から情報を使い易いものにするために，何らかの体系づけを行なっている．この体系づけが分類である．

分類は，情報の検索順を変えるだけであり，元の情報は失われない．言い替えるならば，無秩序であった情報の集まりに対して，一定の基準に基づいた並び順を与えることで秩序を発生させるのが分類である．

Classification and Retreival of NetNews articles using Virtual Newsgroups

Kazufumi Mitani†, Hiroyuki Minami‡, Eiichi Miyamoto†
Faculty of Engineering, Hokkaido University, Nishi 8, Kita 13, Kita-ku, Sapporo 060, Japan
Department of Information and Management Science, Otaru University of Commerce, Midori 3-5-21, Otaru-shi, Hokkaido 047, Japan

2.1　主要な分類法

ここでは，現在用いられている分類法のうち，基本となるものの特長を述べる [2].

- 演繹的分類：全体をある基準に沿っていくつかの部分に分割し，その部分毎に，更に細かな分割を行なう．この種の分類法は階層構造を持つといった特長がある．
  - 意味による分類：図書館で行なわれているデューイの十進分類法など．NetNews のニュースグループもこの分類である．
  - 関係による分類：動植物の分類，家系図などが当たる．
  - 記号による分類：アルファベット順，50 音順に代表される辞書式配列など．最も簡単で，最もよく利用されている分類法である．
- 帰納的分類：統計的な手法によって，要素ごとの類似度を調べ，近いもの同士を集める方法．分類する対象，あるいは対象同士の関係について何らかの数値的データが必要となることが多い．そのため，一般には演繹的な分類よりも複雑である．
  - クラスター分析による分類：要素相互の距離による分類．この場合の「距離」とはどれだけ似ているかという類似度である．
  - ネットワークによる分類:要素間の関連関係による分類．大量のメンバーの総体をきちんと体系化することを諦め，その代わりに個々の要素を関連関係，階層関係，類似性などで相互に結び，ネットワークを作る．有名なものとしてシソーラスがある．
  - 時間による分類：使用頻度による分類．最新の情報，最近利用した情報の方が，

長い間使われなかった情報より再利用されやすいことを利用している．検索に重点を置いた分類で，CPU などのキャッシュと同じ考え方である．ただし，利用順に重要な意味があるため，大勢の人が頻繁に使うような情報には向かず，個人やグループ内での利用に効果がある．

実際，少しの曖昧さもなくすべての要素が分類できる分類法というものは報告されておらず，目的に応じた選択がなされている．

3 仮想ニュースグループによる記事の分類と探索

NetNews 上のニュースの記事は，電子メールの標準規格である RFC822 に制約を加えた RFC1036[3] に従っている．個々の記事にはヘッダ部と本文があり，ヘッダ部には記事の管理や伝達などに必要な情報が，本文には記事の内容が書かれている．

NetNews において，記事の類似度を求めること自体は不可能ではないと思われる．しかし，属性を数値化し類似度を得るための有効な手段が確立されていない以上，計算機によって完全に自動化することは困難である．従って現状では，属性に基づいた数値化を行なう場合，人間の主観に頼らざるを得ない．そこで，ここではヘッダ情報と記事の傾向について，経験的に定めた数値を用いることとして，類似度を考える．

まず，記事には次のような特長があると仮定する．

- 親子関係にある記事は類似度が大きい．
- サブジェクトが同じ記事は類似度が大きい．
- 同じニュースグループの記事は類似度が大きい．
- 加えて，発信された日付が近ければ類似度が大きい．

更に，類似度から距離関係を考え，記事を配置する仮想空間を想定する．距離が近い記事の間には，お互いが引き合う力，つまり引力とでも呼ぶべきものが，関連の無い記事同士には反発する力，つまり斥力が働いているとそれぞれ見なすことで，記事の仮想空間では，関係のある記事は近くに，関係の無い記事は遠くに配置されることになる．

そして，相当数の記事によって構成された空間に対し，自分が欲しい情報を指すようなダミー記事を空間内に新たに加え，ダミー記事が配置された『周辺』の記事を集めて仮想ニュースグループとする．この仮想ニュースグループにアクセスすることにより，記事が実際に投稿されたニュースグループにかかわらず，求めたい情報にアクセスすることが可能となる [1].

3.1 記事空間を用いた分類

記事相互の類似度が与えられている場合，前節で述べた記事空間を作成するには多次元尺度構成法 (MDS) を用いることができる．MDS は類似度を入力として，個体相互間の関係を空間上に示すもので，それぞれ近いものは近くに，遠いものは遠くに配置される．実際の計算過程は固有値問題に帰着することができる．

現在のところ，2 つの記事 a, b の類似度 $e_{a,b}$は，記事のヘッダ部分について

$$e_{a,b} = \begin{cases} \text{References:で参照している} & +1 \\ \text{同じニュースグループである} & +0.5 \\ \text{Subject が同一である} & +1 \\ \text{上がどれか満たされ投稿日の} & \\ \text{開きが一週間以内} & +0.3 \end{cases}$$

と経験的に定めた．なお,"Re" を含んだ関連記事も「Subject が同一である」ものとみなしている．

4 終りに

現在までのところ，類似度を計算するための具体的データを収集しており，インプリメントには至っていない．記事空間上の仮想ニュースグループの構成法を，現実に流通している NetNews に対して計算可能なオーダに引き下げるような工夫を行い，NetNews system への実装と News Reader への対応を試みていく予定である．

参考文献

[1] 佐藤浩史，「大量情報源から得られる情報の分類・選択に関する研究」，北海道大学工学部卒業論文, March, 1994

[2] 野口悠紀雄，「『超』整理法」，中公新書, 1993

[3] Mark R.Horton, 'Standard for USENET Messages' RFC1036, Network Working Group, Dec. 1987