システムソフトウェアと
オペレーティング・システム 77－3
マルチメディア通信と分散処理 87－3
（1998. 2. 26）

個人の視点に基づいた情報分類方法の提案とその実現例

庵 祥子　　鈴木 英明　　砂原 秀樹
NTT ソフトウェア研究所　NTT ソフトウェア研究所　奈良先端科学技術大学院大学

概要

情報取得者に快適な情報取得環境を提供するために「個人の視点に基づいた情報分類方法」について提案する。「個人の視点に基づいた情報分類方法」とは、各情報取得者ごとに分類カテゴリを生成し、情報を分類する方法である。本提案では「閲覧した情報＝興味ある情報」と捉え、閲覧した情報を元に各情報取得者ごとの興味に基づいた分類カテゴリを生成し分類を行なった。また本提案の実現例をプロトタイプシステムとして構築し、実験および評価を行なった。この結果、情報取得者の作成した分類カテゴリの平均 62.5% をシステムが作成した分類カテゴリで再現することができた。

A proposal and its actualization of the information categorizing method based on personal viewpoint

Shoko IHORI　　Hideaki SUZUKI　　Hideki SUNAHARA
NTT Software Laboratories　NTT Software Laboratories　Nara Institute of Science and Technology

Abstract

In this paper, we propose the information categorizing method based on personal viewpoint to provide subscribers with comfortable environment to get information. The information categorizing method based on personal viewpoint is to make categories and classify information for each subscriber. We consider subscribed information to be interesting information. We made categories based on interests of each user from observation which information was subscribed. We also implemented, experimented and evaluated the prototype system of our proposal. We confirmed that 62.5% of categories made by subscribers could be reproduced by our system on average.

1　はじめに

近年、インターネットの発達や普及によって、多くの人々が様々な情報を手軽に取得することが可能になった。特に WWW を利用した情報の収集が盛んに行われている。

このような状況下において情報取得者が自分の興味ある情報を取得したり効率よく再び閲覧するためには、自分自身で自分の視点に基づいた情報分類を行わなければならない。これは、既存のシステムによる分類では、個人の視点に基づいた情報の分類が実現されていないためである。これは既存のシステムによる情報分類が、全ての情報取得者に共通するような分類を行なうことが多いことが原因である。

この問題を解決するために「個人の視点に基づいた情報分類方法」を提案する。この分類方法では、情報取得者の情報閲覧行動から個人の興味を推測し、それに基づいて分類の階層構造や枠組を作成して個人の視点に基づいた分類を実現する。

またこの情報分類の実現例として、WWW 上にプロトタイプシステムを実装する。このシステムでは、情報取得者が特定の WWW サーバに掲載されている情報を閲覧する行動から、情報取得者各個人の視点に基づいた情報分類を実現する。またこのシステムを用いて情報取得者による分類をシステムによる分類でどの程度再現できるかについての実験を行ない評価する。

2　情報分類方法に関する考察

個人の視点に基づいた情報分類方法を提案するにあたり、既存の WWW 上における情報分類について考察を行った。

情報取得者自身で分類する場合は、WWW 上で興味ある情報を閲覧した際に、その興味ある情報へのリンクを自分に分かりやすい階層構造のブックマーク等を作成することによって実現されている。

一方システムが情報分類を行なう場合は、いくつかの方法が試みられている [1]。例えば情報取得者に対してあらかじめ「興味に関するアンケート」を行い、その解答に基づいてシステムが分類の階層構造やノードを作成して分類する方法などがある [2]。しかし、この方法では情報取得者に興味を提示してもらうという余分な労力を必要とするだけでなく、情報取得者の興味の変化に動的に対応することが不可能である。

また情報取得者の情報取得行動をもとにシステム側であらかじめ用意してある分類階層のノードを利用して自動的に分類する方法もある [3] [4]。これはシステム側で自動的に情報の分類を行うが、情報提供者によってあらかじめ分類の階層構造が提供されるので、分類に用いるノードの位置付けが決められている (例えば InternetExproler という分類のノードは Microsoft という分類のノードの下にあるなど)。そのために分類の階層構造そのものに情報取得者個人の興味を反映することができないという問題ある。

このように従来のシステムによる分類は、個々の情報取得者にわかりやすいというよりも、むしろどの情報取得者にも共通するような分類の階層構造やノードを最初に与えることによって行なわれる場合が多い。これらのことから問題点は主に以下の 2 点にまとめられる。

1. 分類の階層構造およびそのノードが情報取得者個人の興味に合致していないこと
2. 分類の階層構造およびそのノードの生成が情報取得者の手を煩わせていること

3 個人の視点に基づいた情報分類システムの提案

第 2 章で述べた問題点を解決し、情報取得者に快適な情報取得環境を提供するために個人の視点に基づいた情報分類システムを提案する。ここでは問題の簡単化のために分類の対象とする情報は、特定のWWWサイトに掲載されている情報とする。また本提案では、分類の階層構造のそれぞれのノードにあたる部分を分類カテゴリと呼ぶ。

本提案では、情報取得者の興味の階層構造の表現と情報分類システムの分類の階層構造の表現を関連づけることで情報取得者の視点に基づく分類を作成する。これにより問題点 (1) の解決をはかる。また、「閲覧する情報＝興味がある情報」と仮定し、閲覧された情報に含まれている keyword 群から自動的に情報取得者の興味を推測し、それに適した分類を行うことを目指す。これにより問題点 (2) の解決をはかる。

問題点 (1) の解決方法を具体的に述べる。情報にはその情報の特徴を表すようないくつかの keyrowd 群が存在する。これを利用して、情報取得者の情報閲覧行動を基にして、情報取得者が興味持つ keyword 群および keywrod を推測し自動的に分類を行なう。分類に用いられた keyword 群および keywrod 同士の関連が深い場合は、これらをまとめる形で上位の分類階層を作成する。これのようにして分類の階層構造と興味の階層構造を関連づけることで、情報取得者に適した階層的な情報分類を行うことを可能にする。すなわち、興味が異なれば、この階層が異なるであろうし、興味が似通っていればこの階層も似ることになる。

さらに本手法では情報を参照する行為が個人の興味を変更するトリガとなるため、興味の表現と関連づけられた階層の表現は、興味の変化に伴って動的に変更されることになるという特徴がある。これにより第 2 章で述べた問題点 (2) を解決する。

以下に分類システムの流れ、WWW サイト側および情報取得者側の keyword の集計方法、それら集計データをもとにした分類カテゴリの生成手法について述べる。

3.1 情報分類の流れ

情報分類の流れは図 1 の通りである。情報の特徴を表す keyword 群はあらかじめ抽出されて情報に添付されているものとする。

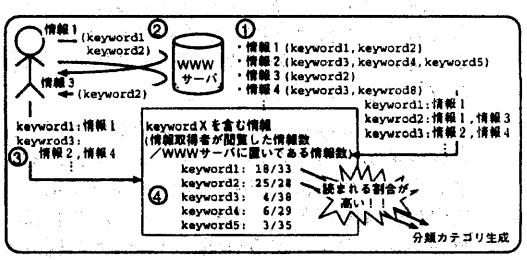

図 1: 情報分類システムの流れ

まず、WWW サイトに掲載された分類の対象となる情報に関して、添付されている keyword 群に基づいて集計する (集計の内容については第 3.2 節で述べる)。その後、情報取得者が WWW サイトに掲載されている情報の中から興味ある

情報を選択して閲覧する。システムは、情報取得者の情報閲覧行動の履歴をとり、閲覧された情報がどのような特徴を持っているかを把握するために、その情報に添付されている keyword 群を集計する。これら2つの集計データをもとに、情報取得者の興味を推測し、分類階層および分類カテゴリを生成する。

上記の分類システムの流れに基づいて、分類システムで必要となる4つの機構についてそれぞれ以下に説明する。

1. WWW サイト上に掲載された情報の keyword の集計
2. 情報取得者による情報閲覧行動の履歴の蓄積
3. 情報取得者が閲覧した情報の keyword の集計
4. keyword の集計データを元にした分類階層および分類カテゴリの自動生成

(1) の機構は、WWW サイトに掲載する際に情報に添付されている keyword 群に関して集計を行う。集計の内容は、「(a) 掲載されている情報と keyword の対応」および「(b) ある keyword に関係する掲載されている情報の個数」である。例えば、今までに情報1～5迄の情報を掲載しており、さらに「keywordA」「keywordB」「keywordC」を持つ「情報1」を掲載する場合、集計は「(a) 情報1: keywordA, keywordB, keywordC」「(b)keywordA:3個」などとなる。

本提案では「閲覧する＝興味がある」であると捉え情報取得者が閲覧した情報をもとにして分類を行うため、(2) の機構で情報取得者の情報閲覧行動の履歴を蓄積する。履歴は情報取得者ごとに蓄積し、集計を行う。

(3) の機構では、情報取得者の情報閲覧行動の履歴を利用して、閲覧した情報にあらかじめ添付されている keyword についての集計を定期的にを行う。集計は「ある keyword に関係する情報取得者が閲覧した情報の個数とその情報 ID」について行う。

(4) の機構では、WWW サイトに掲載されている全ての情報に含まれている keyword と情報取得者が閲覧した全ての情報に含まれていた keyword の比較を行い、各情報取得者に合わせた分類階層および分類カテゴリを自動生成する。これは (2) および (3) の機構で集計したデータを基にして、その keyword を含む分類階層や分類カテゴリを生成するか否かの判断を行う (集計データをもとにした分類カテゴリ生成の判断の詳細については第3.2節で述べる)。これにより、分類カテゴリを個人の興味に近づけることができる。

また、(2) および (3) の集計データがある程度更新されるごとにこの機構を用いて分類カテゴリの再生成を行うことによって情報取得者の興味の動的な変化への対応を可能にする。再生成の際は、集計データに基づいて全て再計算を行う。

3.2 集計データの利用について

ここでは第3.1節の (4) で述べた機構で、どのように集計データを利用して分類カテゴリ生成の判断をするかについて説明する。

分類カテゴリの生成は、情報取得者の興味に基づいて行うため、情報取得者の閲覧した情報の keyword の集計データと WWW サイトに掲載された情報の keyword の集計データを比較した際の数値を基準とする。基準となる数値は、それぞれの keyword を含む情報に対する興味の度合と、2つの keywrod の関連の度合の2つとする。

それぞれの keyword を含む情報に対する興味の度合とは、どれだけ情報取得者がその keyword に興味を示したかを測定するための数値である。keywordA を含む情報に対する興味の度合は、「WWW サイトで提供する keywordA を含む情報の数」と「情報取得者が閲覧した keywrodA を含む情報の数」を比較することによって計算する (式1)。

(式1) keywordA を含む情報に対する興味の度合

$$\text{興味度A} = \frac{\text{情報取得者が閲覧したkeywordAを含む情報の数}}{\text{WWWサイトで公開しているkeywordAを含む情報の数}}$$

すべての keyword に関してこの興味の度合を計算し、相対的に興味の度合の高い keyword について順次分類カテゴリの生成を行う。このようにして分類カテゴリを生成することによって、情報取得者の視点に基づいた情報分類を自動的に行うことを可能にする。

2つの keyword の関連の度合とは、分類の階層化を行なうために用いる keywrod 同士の関連の深さを情報取得者の主観に基づいてはかるためのものである。これは、それぞれの keyword の集合の重なりあったの部分で測定する (図2)。よって関連の度合は「情報取得者が閲覧した keyword A を含む情報の数」と、「情報取得者に閲

覧したkeyword Aかつkeyword Bを含む情報の数」を比較することによって計算する(式2)。さらに関連の度合いが相対的に高いkeywordの組合せについては、単独のkeywordの時の同じく興味の度合の計算を行なう(式3)。このようにして2つのkeywordを含む上位階層の分類カテゴリを生成する。2つ以上のkeywordの関連については、この2つのkeywordの関連の組合せとして扱う。たとえば、keywordA, keywordB, keywordCの関連は、keywordAとkeywordB、keywordAとkeywordC、keywordBとkeywordCの関連として扱う。

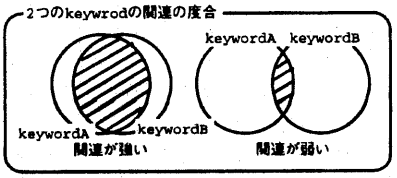

図2: 2つのkeywordの関連の度合

(式2) keywordAのkeywordBに対する関連の度合

$$関連度A(AB) = \frac{情報取得者が閲覧したkeywordAかつkeywordBを含む情報の数}{情報取得者が閲覧したkeywordAを含む情報の数}$$

(式3) keywordAかつBを含む情報に対する興味の度合

$$興味度AB = \frac{情報取得者が閲覧したkeywordAかつkeywordBを含む情報の数}{WWWサイト上のkeywordAかつkeywrodBを含む情報の数}$$

4 プロトタイプの実装および実験評価

第3章に述べた機構の実現に考慮して「個人の視点に基づいた情報分類方法」のプロトタイプシステムの実装を行った。本プロトタイプシステムは、特定のWWWサイトに分類の対象とする情報を情報IDとその情報の特徴を表すkeyword群を添付して掲載し、それを情報取得者が閲覧するという条件のもとで「個人の視点に基づいた情報分類」を実現した。また、本プロトタイプシステムを利用して簡単な評価実験を行なった。

4.1 プロトタイプシステムの設定

本プロトタイプシステムでは、コンピュータやインターネットに関する情報を実験の対象した。情報の特徴を表すkeyword群には、あらかじめ3～6keywordを人間の手作業で抽出し添付しておいたものを利用した。keywordの種類は約250種にわたった。

本プロトタイプシステムではWWWサイトに4つ以上の情報が掲載されているkeywrodについて第3.2節で述べた興味の度合と各keywordの関連について計算した。そして、興味の度合が33%を越えたkeywordについて順次分類カテゴリを生成した。さらに分類カテゴリとして生成されたkeywordについて、2つのkeyword同士の関連の度合が40 %を越えたものについて分類カテゴリの上位階層を生成した。

4.2 実験

評価実験の目的は主に以下の2点である。

1. 情報取得者の興味に適応した分類階層および分類カテゴリが生成されているか
2. 動的に分類カテゴリが更新されているか

実験は、部署内の10名を対象とし12月8日から1月14日の期間に実施した。この間2度に分けてデータ収集とアンケートを実施した。実験期間中にWWWに掲載した情報数は約200、読まれた情報数は延約700であった。実験の内容は特定のWWWサイトに掲載した情報を情報取得者に閲覧してもらい、そのアクセス履歴を用いて個人ごとの視点に基づく分類を行った。

次に実験の流れについて説明する。まず、情報取得者に特定のWWWサイトの情報についてタイトルのみ表示するページで閲覧してもらった。興味もった情報については、そのタイトルをクリックして本文情報も閲覧してもらった。間違えて興味のない情報のタイトルをクリックしてしまった場合は、「興味がなかった」ということを表すボタンをクリックしてもらうことによって、興味を持った情報のみ扱うことを試みた(図3)。

(1)の目的を評価するために、閲覧した特定のWWWサイトの情報について情報取得者自身の判断の分類も行ってもらい、システムの生成した分類との比較を可能にした。なお、この情報取得者自身による分類カテゴリのラベルや階層化は全て自由形式で行った。また情報取得者の判断で分類を行なってもらうために、今回の実験では、システムで生成した各情報取得者の分類カテゴリを情報取得者に表示しない形で行なった。

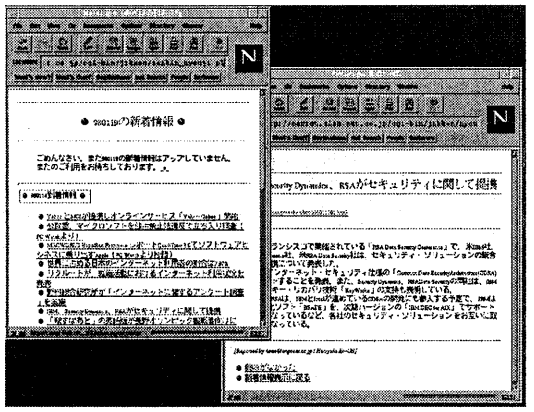

図 3: システムでの情報閲覧

(2) の目的を評価するために、実験期間中 2 回アンケートを取り、各情報取得者の興味や分類カテゴリの変化について調査した。

4.3 評価

評価は、実験の精度をあげるためにある程度の信頼性が保証される実験データに限定して行なった。これは今回の実験データを第 4 章で述べた計算式に当てはめて評価する際に、分母や分子が極端に小さいと実験の精度を下がるためである。具体的には以下の基準で行なった。

- 5 つ以上の情報に含まれる keyword に関する実験データ
- 情報取得者が閲覧した情報の 3 つ以上に含まれる keyword に関する実験データ

この基準にあてはまる実験データをもとに、システムで生成した分類カテゴリと情報取得者が生成した分類カテゴリの比較を行なった。ここでは分類カテゴリのラベルが異なっていても、格納されている情報の 80% 以上がシステムで作成した分類カテゴリに格納されている場合、同一のカテゴリとみなした。さらに情報取得者の作成した分類カテゴリを 1 として、システムで生成された情報取得者と同一の分類カテゴリ数の割合を計算した。今回の実験ではこれを再現率と呼び評価尺度とした。この実験では再現率の分布は以下のようになり、平均は 62.5% であった。

再現率	人数
31%-40%	1
41%-50%	1
51%-60%	2
61%-70%	3
71%-80%	3

情報取得者側では生成されず、システム側で生成された分類カテゴリについては、「意識していなかったけれどもそのようなことに興味を持っていると思う」「そういう分類の仕方もいいと思う」などの意見をアンケートから得られた。これは情報取得者の潜在的な分類カテゴリをシステム側の生成した分類ノードによって引き出すことができたと考えられる。

図 4 は実験結果で得られた情報取得者に基づいた分類カテゴリの一例である。この例では、情報取得者側の分類カテゴリで表されている組織関連、Microsoft 関連、Netscape 関連、電子商取引関連、Java 関連のカテゴリとそのその下位カテゴリの大部分が、システム側にも分類カテゴリとして生成されていることがわかる。

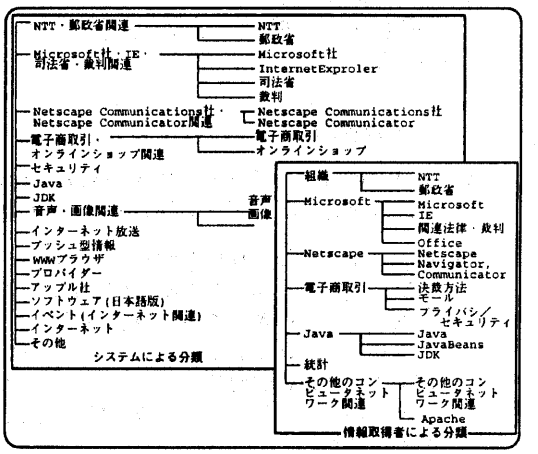

図 4: 実験結果の例

システム側で作成した分類カテゴリと情報取得者が作成した分類カテゴリの傾向について述べる。分類カテゴリが一致する傾向にあったのは、「Microsoft」,「Java」,「NetscapeCommunicator」などの固有名詞がラベルとされている分類カテゴリである。これらの固有名詞はシステム側の keyword になりやすく、かつ情報

取得者側もそれらのkeywordの出現する情報をその固有名詞をラベルとした分類カテゴリに分類したからであると思われる。

これに対して、情報取得者が「コンピュータ」や「ネットワーク関連」等のラベルをつけた分類カテゴリについては、システム側でうまく再現することができなかった。この理由としては、これらの分類カテゴリは、情報取得者ごとに想像する範囲が異なること、また非常に範囲の広い情報が集められており、システム側と分類の規模の一致がうまくはかれなかった場合があったことなどが考えられる。また、「趣味」などのラベルのついた分類カテゴリも、情報取得者ごとに多様に渡っており、システム側だけでの分類は不可能であった。

これら範囲の広い情報が集められている分類カテゴリは、そこに所属する情報の個数が多くなるにつれて細分化される傾向にあった。細分化された分類カテゴリについては、システム側の分類との一致が見られる場合が多かった。

また、興味ある情報をより多く閲覧している情報取得者は分類カテゴリの再現率が70%以上であった。これは閲覧した情報が多くなることによって情報取得者とシステム双方の分類カテゴリの詳細化が一段落し、安定したためであると思われる。

5 今後の課題

本実験で得られたデータについて考察した結果、今後の課題として以下の4点が挙げられた。

- keywordの予測の傾向をつかむ
  今回の実験で、情報のタイトルに出現していないのに着目されているkeywordがいくつか見受けられた。このようなタイトルに直接出てこないが想像されるkeywordについての傾向を分析するための仕組みを実験に組み込む。

- 有効なタイトルの自動化
  今回の実験から、似たような情報でもタイトルによって閲覧されたりされなかったりすることが分かった。これはタイトルの付け方によって、閲覧される割合が変わってくることを表している。そこで、閲覧される割合の高い有効なタイトルの付け方について解析を行ない、自動化を目指す。

- 情報取得者の比較
  この分類システムで生成される分類カテゴリは情報取得者ごとの特徴を表すデータとなる。このデータを利用して、興味の似通った情報取得者との出会いを支援するシステムの構築を検討する。

- 認知科学的見地からの考察
  情報取得者による分類カテゴリの生成とシステム分類カテゴリの生成についてて認知科学的な見地からの考察を行ない、より両者を近付けるための法則について検討を行なう。

6 結論

プロトタイプシステムの実験から、情報取得者が分類に用いた分類カテゴリの平均62.5%をシステムの分類カテゴリとして生成することができた。これにより、個人の興味に適した情報分類が可能となり、本提案の「個人の視点に基づいた情報分類方法」は有効であることが分かった。

しかしながら、今回の実験は実験期間が短かったため、充分な実験データがとれなかったたり、興味のある情報が少なかった被験者からは、システムによる階層化を行なうことがかったなどの問題点があった。

今後これらを改善してさらに実験を行ない、今後の課題で述べたことについても研究を勧めていく予定である。

参考文献

[1] 森田昌宏：情報フィルタリングに関する研究動向, JAIST Research Report, IS-RR-93-9I(1993)

[2] 萩野浩明、門林理恵子、塚本昌彦、西尾章治郎: 推論システムを利用したメッセージ管理システムにおける分類・検索機構, 人工知能学会全国大会(第10回)論文集 pages557-560

[3] ソムヌック サグアントラクーン、寺田努、塚本昌彦、西尾章治郎、三浦康史、松浦聡、今中武: 放送型データのユーザ適応型分類・選択手法, 情報処理学会研究報告 DPS(1997.11)

[4] 野美山浩、紺谷精一、渡辺日出雄、串間和彦、堤泰治郎: 個人適応型情報検索システム, 情報学基礎 42-8,(1996.7.26)