システムソフトウェアと
オペレーティング・システム 77－1
マルチメディア通信と分散処理 87－1
（1998. 2. 26）

利用履歴に基づくコンテンツ分類手法の検討

坂本 啓，金井 敦

NTT ソフトウェア研究所
日本電信電話株式会社

WWW が提供する情報空間には現在多数の情報が存在するが、実際に情報を取得しようとした時に大量の無関係な情報の中に本当に意味のある情報が埋もれてしまい立ち往生するという事態がしばしば発生する。このような問題は、単に情報が大量にあるというだけでははく、それらが構造化されていないことにも大きな原因がある。本稿では他者の利用痕跡を活用して情報の構造化を行なうために、コンテンツの利用履歴から、特定の利用者集団により主に選択されるようなリソース群を抽出する手法について提案する。さらに実験を通じて提案手法により利用履歴から情報の分類・構造化が可能であることを確認した結果について報告する。

Structurizing Technique of Resources Based on Browsing Behavior

Akira Sakamoto, Atushi Kanai
NTT Software Laboratries
NTT
3-9-11 Midori-cho Musashino-shi Tokyo 180-0012 Japan

When someone gets information by WWW, he/she often experience that it is very hard to discover useful infomation because it is buried in a large amount of information. This problem is caused that not only there are huge resources in WWW but resources are not structurized. In this paper, we use records of access to resources by user for structurizing them. We propose a technique that a resource group chiefly selected by a specific user group is extracted from the access records. In addition, we made an experiment to verify the effectiveness of a proposed technique. It was confirmed to be able to classify and to structurize resources from the access recoeds as a result.

1 背景

インターネットの利用は拡大の一途を辿り、これをプラットフォームとした商取り引きになどにも利用されるようになってきた。これに伴い公開されているWebサイトも莫大な数にのぼっている。このような状況の中でWWWが提供する情報空間中に有効な情報が多数存在し使えるものとなってきた反面、実際に情報を取得しようとした時に大量の無関係な情報の中に本当に意味のある情報が埋もれてしまい立ち往生するという事態がしばしば発生する。このような問題は、単に情報が大量にあるというだけではなく、それらが構造化されていないことにも大きな原因がある。情報の構造化には文書から抽出したキーワードから特徴ベクトルを形成し、文書空間中に埋め込むといったアプローチがとられることが多いが、ここでは少し別の角度からこの問題について取り扱う。

我々が日常生活をおくる中で、他者の行動の痕跡から環境の使われ方を暗黙裏に読みとることでその環境に合わせた行動をそれほど苦労することなくとっていることは案外多いのではないだろうか。例えば、極端な話だが人気の無い山中で不安に思いながら歩いている時、ぬかるみに足跡を発見して大いに安心するといった場合を考えてみる。これは、現在の進行方向が正しい道だという自分の判断に基づいて行動しながらも、他者の足跡という環境の利用痕跡を見い出すことで、その環境の利用方法についての自分の判断が正しいという保証を外側から得ることができ、結果的にもともとの判断を強化している、という例になっているだろう。つまり対象に対する他者の振舞の痕跡というのは、対象に対する価値判断の有力な根拠となりうる。

これに対しWWWが形成する情報空間のブラウジングは、情報空間とそれに対峙する私しかいないという意味で非常に個人的な行為であることがわかる。その情報空間を実は非常に多数の人間が通過し、その中で情報を検索・探索を行ない、コンテンツが利用されているにもかかわらずである。これはその時対峙している情報の価値を判断する際、その情報それ自体とその情報に至るまでに経てきた個人的なコンテキストだけしか利用できない、ということである。

以上の状況を踏まえて本稿では、他者によるコンテンツへの利用を痕跡として積極的に残し、それを利用者にフィードバックすることを考える。具体的にはコンテンツの利用履歴を使ってコンテンツの関係を見い出し、コンテンツ分類を行なうことを通して、情報空間を構造化することを試みる。これは個々人の明示的な興味の提示によりコンテンツ選択を行い積極的に提示したり、コメント付けなどにより他者が積極的に介在しコンテンツに重み付けをしたりといった、いわゆるソーシャルフィルタリング[1]で行なわれる手法とはアプローチを異にする。コンテンツの関係性という形で定着された他者の利用痕跡は新たにその情報空間を利用する者にフィードバックされることで利用者のブラウジング行動を支援するために用いる。しかしフィードバックされた関係性から何を読みとり、その上でどういうコンテンツ利用を行なうかは最終的に個々人の判断に委ねられるのが理想であろう。つまり本研究の目標は、利用者のブラウジング行動を(肯定的な意味で)消極的に支援することにある。

以下第2章でここで取り扱う問題のより詳細な定式化を行ない、第3章でその問題に対するアプローチをアルゴリズムの形で提示する。さらにこのアルゴリズムにより利用履歴からどの程度コンテンツ分類が可能かを実際に確認するために、コンテンツとしてCGアート作品を利用した実験を行なった。実験概要とその結果を第4章に示す。最後にまとめと今後の課題を述べる。

2 問題の定式化

通常WWWとは「ハイパーリンクで相互に接続されたページの集合体」として理解される。利用者は、WWWが提供するページ集合体としての情報空間の中を、ページ移動をしながら個々のページの内容を観賞することを通して、その利用目的を達する。したがってWWW情報空間中の個々の利用者による利用軌跡は、グラフノード上の遷移として記述できる。

話を単純にするために、ここでは利用者の嗜好以外の属性に顕著な差のないn個のリソースが互いにハイパーリンクにより接続されており、全体として1つの情報空間が構成されている場合を考える。この情報空間は構造として有向グラフをなし、任意のリソースからリンクの張られているリソースへは自由に移動可能なものとする。ここでリソース集合を$R := \{r_i | 1 \leq i \leq n\}$、リソース間のハイパーリンク集合を$E := \{e_{ij} | r_i から r_j へのリンク\}$とするとき、コンテンツ$C$はノードとなるリソースの集合$R$と辺となるハイパーリンク集合$E$からなる有向グラフ$C := \langle R, E \rangle$により表現できる。

用意したコンテンツに対しm人の利用者が訪れたとする。利用者は各自の嗜好に基づき適当なリソースにアクセスする。利用者jによりアクセスされた一連のリソースをs_jとし、これをセッションと呼ぶ。セッションはCの部分グラフとなるが、今回は問題を簡単にするためリンク関係を意図的に無視しアクセスされたリソースにのみ着目する($s_j \subseteq R$)。

ここで利用者は、その利用特性(嗜好)により M ($\ll m$) 個のクラスに分類でき、全ての利用者(上記の m 以外の潜在的利用者を含む)はこの M 個のクラスのいずれかに属しているものと仮定する。このとき利用者は各々所属するクラスに特徴的な利用属性に従い特徴的なリソース選択を行なうこと、言い換えればあるクラスに属する利用者集団により主に選択されるようなリソース群が存在すること、を期待する。

以上のような状況下で、与えられた l 個のセッション

$$S := \{s_j | s_j \subseteq R, 1 \leq j \leq l\}$$

に基づいて利用特性のクラスの数 M と各クラスに属する利用者により主に選択される特徴的なリソースを求めることがここでの問題である。すなわち、セッション集合 S 中に頻出するリソースの組み合わせ全体の集合を δ、δ の部分集合全体の集合を $\mathcal{D}$ とし、さらに $\mathcal{D}$ の要素

$$D := \left\{ d_k \middle| 1 \leq k \leq M,\ d_k \in \delta,\ \bigcup_k d_k \subseteq R \right\}$$

に対して

$$f(D) = \sum_{i<j}^{M} | d_i \cap d_j | \quad : \text{重複数}$$

$$g(D) = \sum_{i}^{M} | d_i | \quad : \text{被覆数}$$

ただし $|\cdot|$ は集合の要素数

を規定するとき、$\mathcal{D}$ から下記制約を満たす D を求めるわけである。

制約1. 重複数ができるだけ小さくなるような $D(\in \mathcal{D})$ を求める。

制約2. 被覆数ができるだけ大きくなるような $D(\in \mathcal{D})$ を求める。

制約3. M ができるだけ小さくなるような $D(\in \mathcal{D})$ を求める。

ここで得られた D を**パターン**、パターンを構成する個々のリソース群 d_k を**特徴リソース**と呼ぶ。

以上の定式化の下で、

1. セッションの集合からどのようにすればパターンが抽出できるか
2. 雑多な利用履歴であるセッションの集合から本当にパターンが抽出できるか

について、考察・確認することが本稿の主題である。

3 パターン抽出アルゴリズム

前章で述べたようにパターンは多数の利用履歴データから抽出される。大量に蓄積された生データから意味のある知見を抽出する手法としてデータマイニング [5] が注目を集めているが、データマイニングにおける主要な技術として相関ルールの抽出がある。相関ルールに関する詳細は [6] に譲るが、ここでは前章に示した問題を相関ルール抽出の拡張問題と位置付け、相関ルール抽出アルゴリズムをベースにしたパターン抽出アルゴリズムについて述べる。

パターン抽出アルゴリズム

1. 与えられたセッション集合中に頻出するリソースの組合せ(部分リソース集合)を全て求める。ここで得られた個々の部分リソース集合を特徴リソースと呼ぶ。
2. 特徴リソース集合に基づき、第2章であげた制約1. を満たす組み合わせを求める。ここで得られたパターンの候補となる組合せの集合を探索集合と呼ぶ。
3. 探索集合から制約2., 3. を満たす組合せを抽出する。

第一ステップは、相関ルール抽出の第一ステップと全く等価と見なせるから具体的な特徴リソースの導出には、相関ルール抽出の代表的アルゴリズムであるアプリオリ [6] を特に修正することなく利用することができる。

セッション集合の中で注目している部分リソース集合を含むセッションの割合をその部分リソース集合の“支持度”と呼ぶ。このときアプリオリアルゴリズムは、分析者により与えられた最小指示度を満足するような部分リソース集合を効率的に全て見い出すものである。支持度は、注目している部分リソース集合がセッション集合の中で頻出するリソースの組合せであることを保証する指標であり、ここで保証されたリソースの組合せはあるクラスに属する利用者集団により主に選択されるようなリソース群と解釈することができる。したがってアプリオリを用いて求められた部分リソース集合は特徴リソースであるといえる。以下特徴リソースを求めるアルゴリズムの詳細を示す。

最小支持度を満足する大きさ k の特徴リソース全体の集合(特徴リソース集合)を L_k, 大きさ k の特徴リソース集合の候補(候補集合)を C_k とすると、大きさ $k \geq 2$ の場合の処理は次のようになる。

1. 大きさ $(k-1)$ の特徴リソース集合 L_{k-1} から大きさ k の候補集合 C_k を作成する。

2. C_k の各要素についてセッション集合に基づいて支持度を求める。

3. C_k から最小支持度を満足しないものを取り除く。これを大きさ k の特徴リソース集合 L_k とする。

4. 以上の処理を新しい特徴リソース集合が空になるまで繰り返す。

第二ステップでは、第一ステップで求めた特徴リソースを組み合わせて制約 1. を満足するもの最終的に残すわけだが、その組合せの数は莫大なもの[1]となり総当たり的に求めるには限界がある。

そこで一気に制約 1. を満たす探索集合を求めるのではなく、あらかじめ $f(D)$ に上限を与えておくことで探索しなければならない組み合わせの数を縮小することを考える。すなわち全リソースのうち複数の特徴リソースに属しているリソースの割合を"重複度"と呼ぶとき、分析者により与えられた最大重複度を満足するような特徴リソースの組合せを見い出すようにする。この結果、組み合わせの数は小さくなるが、得られる探索集合は若干大きなものとなる。

以下探索集合を求めるアルゴリズムを示す。

対象となる特徴リソースを $d_i\ (1 \leq i \leq N)$、k 個の特徴リソースの組み合わせとして得られる候補パターンからなる候補パターン集合を $\mathcal{P}_k$, 実際の探索集合を $\mathcal{D}_k$ とする。$\mathcal{D}_0 = \phi$ を初期状態として、$k \geq 1$ の場合の処理は次のようになる。

1. 大きさ $(k-1)$ の探索集合 $\mathcal{D}_{k-1}$ から大きさ k の候補パターン集合 $\mathcal{P}_k$ を作成する。

2. $\mathcal{P}_k$ の各要素についてそれぞれ重複度を求める。

3. $\mathcal{P}_k$ から最大重複度を満足しない候補パターンを取り除く。残った候補パターンの集合を大きさ k の探索集合 $\mathcal{D}_k$ とする。

4. 1.〜3. の処理を 1. で候補パターン集合が作れなくまで繰り返す。

また候補集合の構成 (ステップ 1) で用いる具体的なアルゴリズムを示す。

1. ある $D_{k-1}\ (\in \mathcal{D}_{k-1})$ に含まれる特徴リソースの中で最大の添数を求め、これを k_{max} とする。

2. $k_{max} < N$ であるとき、$k_{max}+1 < x \leq N$ を満たす全ての x について、それぞれ $\{d_x\} \cup D_{k-1}$ を作成し、$\mathcal{P}_k$ の要素として加える。

3. 以上を $\mathcal{D}_{k-1}$ の全ての要素に対して行なう。

最後に第三ステップは探索集合から $g(D)$ が最大のものを選び、さらにそこから $f(D), M$ を最小とするような組合せを抽出するだけである。したがって第一，第二ステップに比べると計算量的にはほとんど問題にならない。

4 実験

前節で示したアルゴリズムによりどの程度興味深いパターンが抽出できるか、を確認するために、実際に一般に公開しているサーバ [4] 上に実験環境を構築して実験を行なった。

4.1 実験環境

実験で用いた Web サーバは oneZero Project[2]が運用するものである。このサーバは CG 作家の任意団体「デジタル・イメージ」の協力を得て、彼らの CG 作品や展覧会の情報などデジタルアートに関する情報提供を行なっている。

利用者の嗜好にのみ依存し、それ以外の属性に顕著な差のないリソースとして、ここではそれぞれ異なる作家により描かれた 35 枚の静止画を用いた。このリソースをノードとする双方向の完全有向グラフがここで実験対象となるコンテンツである。実験に用いた具体的なページ構成を図.1 に示す。実験環境は 2 つのブラウザウィンドウから構成され、1 つはリソースのサムネイルが表形式で表示されたメニューウィンドウ、もう 1 つは利用者が注目したリソースが拡大表示さた作品ウィンドウである。この実験環境へは、トップページから他のメニュー同様にアクセスできるようになっており、そのリンクを辿るとまずメニューページが表れるようになっている。

メニューウィンドウと作品ウィンドウを分けたために、任意のリソースが表示されている状態から他のリソースへ表示の切替えを行なう際、操作上の相違はない。このためコンテンツ上の選択行動は、利用者の嗜好のみに左右されることが期待される。

以上のような環境を構築し、一定期間外部に向けて公開して一般利用者に自由に利用してもらい、データ収集を行った。

4.2 結果

利用者の利用履歴は Web サーバのアクセスログとして蓄積される。蓄積したアクセスログから実験環境に関連するリソースに対するものを抽出し、同

[1] 可能な特徴リソースの数が N であるとき $\sum_{i=1}^{N} {}_N\mathrm{C}_i$ 通りの組み合わせが存在する。

[2] oneZero Project は、株式会社シフカと NTT ソフトウェア研究所との共同研究プロジェクトである。

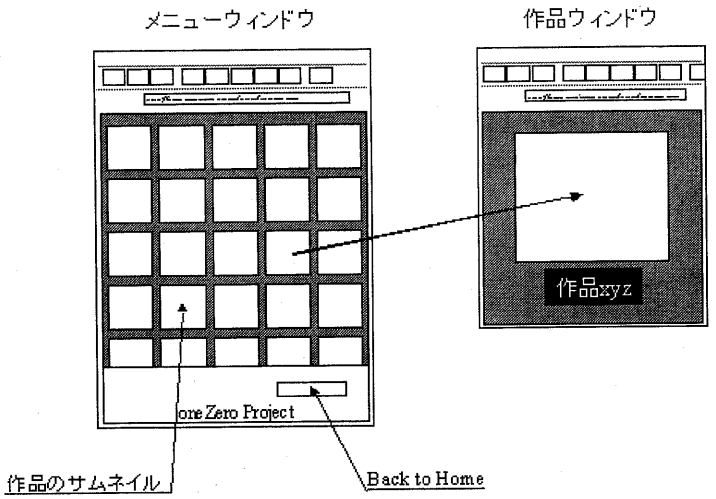

図 1: 実験環境のページ構成

一ホストからのアクセスを時系列にしたがってまとめた形 (セッション形式 [2] ここで得られるセッション形式が、第2章でセッションとしたものに相当する。なおアクセスログからの情報では厳密には個人の利用履歴を特定することはできないが、ここでは以下の仮定を設けることで、特定ホストからの一連のアクセス履歴を個人の利用履歴と見なしている。

仮定 1 : ホストとユーザは 1 対 1 で対応する。

仮定 2 : 同一時間中に 2 つ以上のホストからのセッションが重複することはない。

仮定 3 : 同一セッション内でアクセス間隔が 30 分以上空くことはない。

得られたセッションに対し、第3章で示したアルゴリズムを適用することによってパターンを求めた。

実験期間は 1997 年 12 月 28 日〜 1998 年 1 月 20 日の約 3 週間であり、この間約 300 ホストから実験環境に対してアクセスがあった。この中で全てのリソースへの参照があった場合を除いた 280 セッションを今回分析の対象とした。表.1 に個人および作品毎に見た平均参照数, 最小・最大参照数および分散を示す。

表 1: 基礎統計量

	平均参照数	最小参照数	最大参照数	分散
個人	3.9	2	25	11.1
作品	31.5	12	72	242.8

このようなセッション集合に対して、最小支持度 0.017 として特徴リソースを求めた。この結果大きさ 5 まで合計 491 個の特徴リソースが得られたが、このうちより大きな特徴リソースの部分となっているものを除きかつ大きさ 4 以上である 26 個を最終的な特徴リソースとして残した。

さらに得られた特徴リソース集合に対して、最大重複度 0.14 として探索集合を生成し、最終的に被覆数 17, 重複数 4, 特徴リソース数 7 となる 3 つのパターンを得た。その中の一つは以下のようであり、各特徴リソースの関連を図示したものを図.2 に示した。なおここでは作品にユニークに付加した番号で個々の作品を表している。

d_1 = {3,15,18,20,35}
d_2 = {3,20,27,28,35}
d_3 = {19,20,29,33,35}
d_4 = {2,20,33,35}
d_5 = {3,10,21,25}
d_6 = {12,20,33,35}
d_7 = {20,30,31,35}

得られたパターン

4.3 考察

今回採用した手法と多変量解析などで主に用いられるクラスタリング手法との最大の相違は、いわゆるクラスタリングの場合各リソースは必ず 1 つのクラスタにのみ所属し、重複がないのに対し、提案

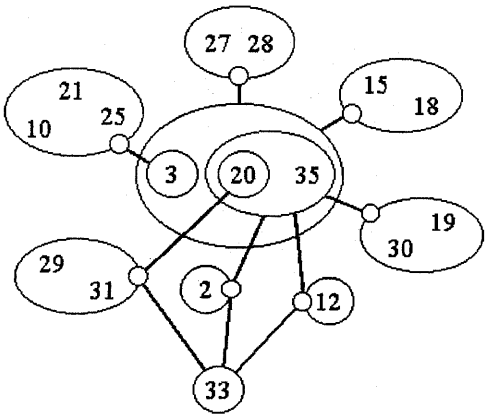

図 2: 結果的に得られたパターン

手法では複数の特徴リソース (クラスタ) に重複して所属することを許している点にある。実際図.2で得られたパターンを見てみると、3,20,33,35 の4つのリソースが複数の特徴リソースで重複している。

個々の特徴リソースは、それぞれ利用特性により典型的な選択行動を引き起こすようなリソース群と解釈される。したがって複数の特徴リソースに重複して所属するリソースは多重の 意味を持っており、異なるクラスに属する利用者毎に違う価値を持って立ち表れることになる。ここでは利用特性の判断基準として"嗜好" のような絶対的かつ一元的な尺度を作れない指標を用いている。このような場合、同じ対象に対して同じように肯定的な評価を下していても理由が異なる場合があるのはよくあることであり、利用のコンテキストの相違により意味付けが違ってくるリソースが存在するのはむしろ自然なことである。提案手法はその点を直接敵に抽出できているものと考えられる。

5 まとめおよび今後の課題

本稿では、コンテンツの利用履歴を使ってパターンを抽出する手法を提案した。さらに実験を通じて提案手法により、利用履歴からどの程度コンテンツ分類が可能かを確認した。

今回は、パターンの抽出問題を相関ルール抽出の拡張問題として捉えることで、データマイニング的な手法を検討した。しかし効率面から見ると、特に第二ステップで用いた、最大重複度を満足する特徴リソースの組合せの構成法についてはまだまだ改良の余地がある。また得られるパターンとサンプル数の関係についても未検討である。サンプル数が増加していったとき、あるパターンへ漸近的に収束していくものなのか、それともサンプルの個別性がパターンに直接反映されてしまうのか、さらに収束する場合にどの程度のサンプル数があれば十分なのかといった点については未検討である。より効率的なアルゴリズムと得られるパターンの安定性については今後の課題としたい。

また今回の定式化ではリソースに予め付加されているはずの関係や利用履歴の順序関係については検討対象から外してある。WWW の利用履歴を対象とする場合、リソース間の時系列的な関係という制約だけでなく、コンテンツにあらかじめハイパーリンク構造がありその上での状態遷移であるというもともとリソース間に存在する制約加わるため、定式化が複雑になる。このような制約が既にあるリソース間の利用履歴からの関係性抽出についても今後の課題である。

さらに、今回は対象リソースとして CG アートの作品という小さな粒度のものを用い、1サイトのアクセスログから利用履歴を抽出したが、ここでの手法を最終的に Web サイト間の関係性にまで広げていきたいと考えている。Web サイト間の関係性を議論する場合、最大の問題点はサイトをまたがる横断的な利用履歴の取得が難しいということである。この問題はプロキシ上で監視したりクライアント側にモニタリング機構 [3] を組み込むなどの方法で技術的には解決可能である。本稿の延長線として Web サイト間の関係性抽出について検討を続けていくつもりである。

参考文献

[1] 森田昌宏, 速水治夫, 情報フィルタリングシステム, 情報処理学会誌, Vol.37, No.8, PP.751–758, 1996.8.

[2] 坂本 泰久, 岸 晃司, ユーザアクションにもとづく Web サーバアクセス履歴の分析, 情報処理学会シンポジウム Interaction 97, 1997.2.

[3] 坂本 泰久 他, クライアント監視方式による WWW サービス効果測定, マルチメディア, 分散, 協調とモバイル (DiCoMo) ワークショップ論文集, pp.269–274, 1997.7.

[4] oneZero, http://onezero.sl.cae.ntt.co.jp/

[5] 特集「大規模データベースからの知識獲得」, 沼尾 雅之編, 人口知能学会誌, Vol.12, No.4, pp.496–549, 1997.

[6] Agrawal, A., Srikant, R.: Fast Algorithms for Mining Association Rules, Proceedings of VLDB, pp.487–499, 1994.