J-003

# 情報整理のための手動クラスタリングによる人の興味の抽出

## Identification of Human Interests with Manual Clustering for Information Organization

樋口 賢治 † 原田 史子 ‡ 島川 博光 ‡
Kenji Higuchi Fumiko Harada Hiromitsu Simakawa

## 1. はじめに

パーソナルコンピュータやブロードバンド環境構築の低価格化が進んでいる．それに伴い，人々がパーソナルコンピュータに触れ，インターネットに接続する機会は多くなった．人が閲覧する情報量は日々増加し，有益な情報が大量の情報に埋もれるようになった．このことから，閲覧，収集した情報の検索が困難になり，情報を再度活用する機会が失われている．そこで，人が閲覧，収集した情報の適切な整理を支援する手法を開発する必要性がある．

人は，情報に対して何らかの興味や目的があるためにその情報を収集すると考えられる．つまり興味に基づき情報を整理することができれば，それは情報を活用する目的別に整理できることになる．そのためには，人の興味を定量的に抽出できなければならない．

本論文では，人が収集した情報の適切な整理を支援するために，人が収集した情報の履歴から人の興味を抽出する手法を提案する．まず，人の興味が多次元空間上のベクトルとして表現できると仮定する．同様にして，人の収集する情報もその多次元空間上のベクトルとして表現できる．このとき，人がある興味に基づいて収集した情報のベクトルは，その興味のベクトル付近に集まると考えられる．よって，多次元空間上の情報のベクトルを興味毎に分類できれば，人の興味を多次元空間上のベクトルとして求めることが可能になる．本論文で提案する手法では，情報のベクトルの分類をするために，多次元空間上のベクトルを主成分分析法を用いて二次元平面上で表現し，情報収集者自身が手動でベクトルをクラスタリングする．そして，得られた各クラスタの重心を興味のベクトルとして抽出する．

興味のベクトルの有用性，手動によるクラスタリングの妥当性の検証を行った．アンケートにおいて，被験者は抽出されたすべての興味のベクトルのうち，81.96 % の興味のベクトルに対してその興味を連想できると回答した．抽出された興味のベクトルは，収集された情報の検索において3ヶ月後も有用であることがわかった．また，得られたクラスタから，手動によるクラスタリングは情報を人の興味ごとに分類することに適していることがわかった．

## 2. 氾濫する情報の整理

### 2.1 情報の氾濫

近年，パーソナルコンピュータの高性能化に伴う記憶媒体の大容量化，インターネット回線の高速化が進んでいる．パーソナルコンピュータ，インターネットの普及率は年々上昇している [1]．人々が業務以外でパーソナルコンピュータに触れる機会が増え，個人が多様かつ大量の情報を高速に取得することがきわめて容易になった．また，街を歩けば，街頭で配られるポケットティッシュやポスター，電車の吊り広告や，ビルの壁や飛行船，巨大スクリーンに映し出されたコマーシャルなど，現実世界において情報を伝える媒体は多様化している．個人の触れる情報量は日々増加し続け，人間の認知能力を超えて認知限界に達している．その結果，情報リテラシの低い人々は閲覧・収集した情報を適切に整理することができず，また情報の利活用も適切に行えない．そこで，情報の適切な整理を支援する手法を開発する必要がある．

### 2.2 興味による情報整理

人はある興味に基づいて情報を収集すると考えられる．また，人はある目的について収集した情報という観点から情報の検索を行う．ある目的とは，情報を収集したときの興味と考えられる．その興味を情報整理の基準とすることで，興味に基づいた情報の整理および検索が可能になると考えられる．

情報整理の基準に興味を用いると，ユーザにとって情報の収集目的が明確になり，収集した情報に収集の目的付けが行われることになる．よって，ある目的に関する情報の検索が効率化され，情報の利活用を促進することが期待できる．

情報整理のための基準としてユーザの興味を使用する手法が望まれる．またその手法の実現のためにはユーザの興味を定量的に表現する手段が必要不可欠である．

## 3. 情報整理のための興味ベクトル抽出

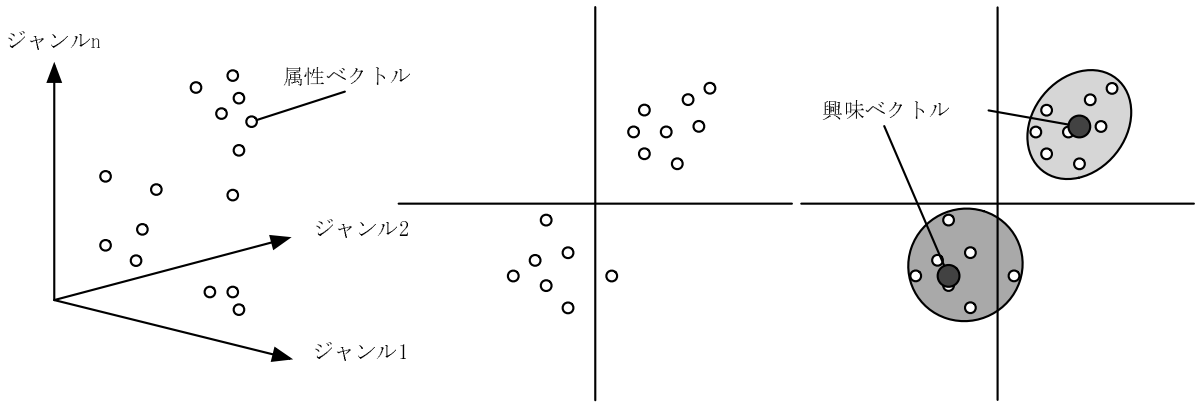
### 3.1 興味と情報のベクトル化

ジャンルを軸とする多次元空間を用意する．ジャンルとはユーザが収集する情報の持つ属性を構成するものである．ジャンルが $n$ 種類あれば，この多次元空間は $n$ 次元のベクトル空間ということになる．ここで，あるユーザのある興味がこの多次元空間上のベクトルで表現できると仮定する．このベクトル化されたユーザの興味を，本論文では興味ベクトルとよぶ．ユーザの興味は同時に複数個存在することが考えられるので，興味ベクトルも同時に複数個存在することになる．

同様にして，ユーザが収集する情報も，多次元空間上のベクトルとして表現できる．多次元空間上のベクトルとして表現した情報の持つ特性を，情報の属性ベクトルとよぶ．

ユーザがある興味に基づいて情報を収集したとき，収集した情報の属性ベクトルは，その興味に対応する興味ベクトルの近くに存在すると考えられる．すると，その興味ベクトルの近くには，その興味に基づいて収集した

†立命館大学大学院 理工学研究科
‡立命館大学 情報理工学部




(a) 多次元空間上の属性ベクトル　(b) 主成分分析法の適用　(c) 興味ベクトルの同定

図 1: 興味ベクトルの同定

情報の属性ベクトルが群として集中することになる．また，ユーザの興味ベクトルは興味の数だけ存在するので，それぞれの興味ベクトルの近くに情報の属性ベクトル群が形成されると考えられる．

このことから，ユーザが収集した情報の属性ベクトル群のもっとも密度の高いところ，つまり，重心付近にユーザの興味ベクトルが存在すると考えられる．すなわち，ユーザが収集した情報の属性ベクトルを興味ごとにクラスタリングすることで，ユーザの興味を興味ベクトルという定量化された形で求めることができる．

### 3.2 手動クラスタリングによる興味の抽出

ユーザが収集した情報には，何らかの方法で属性ベクトルが設定されているものとする．例えば，情報作成者が主観で設定するといったようなものである．このとき，図 1 (a) のように，ユーザが収集した情報の属性ベクトルは多次元空間上に配置できるので，これをクラスタリングすることで興味ベクトルを求めることになる．しかし，興味は個人によって異なるので，機械的な既存のクラスタリング手法を適用するのは不適切であると考えられる．ここで，代表的な二つのクラスタリング手法を適用する例を挙げる．

まず，階層的クラスタリング手法を適用する場合を考える．例えば最短距離法 [2] の場合，クラスタリングの閾値として属性ベクトル間の距離を用いることになる．しかし，人により，また興味により，クラスタの広がりは変わってくると考えられるので，同一の閾値をもとにクラスタリングを行っても，期待通りの結果が得られない可能性がある．

次に，非階層的クラスタリング手法，たとえば $k$ -means 法 [2] を適用する場合を考える． $k$ -means 法の場合，重心すなわち興味の数をもとにクラスタリングを行うことになる．ユーザ自身がいくつの興味をもとに情報収集を行ったと認識しているとしても，収集した情報の中にはユーザ自身が気づかない，潜在的な興味をもとに収集したものも含まれている可能性がある．すると，興味の数が一致しないことになるため，正常にクラスタリングを行えないことが考えられる．

本研究では，ユーザの期待通りの結果を得るためには，ユーザがクラスタリングをユーザ自身の主観に基づいて行う方が確実であると考える．しかし，多次元空間を正確に認識し，クラスタリングを行うのは困難である．システムは何らかの方法で多次元空間を人間に認識できる形に変換してユーザに提示する必要がある．ユーザが手動でクラスタリングを行えるように多次元空間上の属性ベクトルに対して主成分分析法 [3] を適用し，小さな次元の空間上でこれら属性ベクトルを表現する．本研究では図 1 (b) のように属性ベクトルを二次元化する．多次元空間上に配置された属性ベクトルを二次元平面上で表現することで，本来の属性ベクトル同士の位置関係が崩れてしまうおそれがある．そのため，本研究で用いる主成分分析では，もっとも寄与率の高い第一主成分と，次に寄与率の高い第二主成分をそれぞれ抽出し，これらを二次元平面の軸とする．属性ベクトルを二次元平面上で表現することでユーザ自身によるクラスタリングを可能にする．ユーザは図 1 (c) のように手動でクラスタリングを行い，システムは得られたクラスタの重心を同定する．そして，その重心をユーザの興味ベクトルとする．ひとつの属性ベクトルが複数のクラスタに属することは機械的なクラスタリング手法では困難であるが，手動クラスタリングでは容易になるため，柔軟な情報整理が可能になる．

## 4. 実験

### 4.1 実験概要

本実験では，本手法の妥当性および抽出された興味ベクトルの情報整理における有用性について検証する．

実験は，二度にわたり実施された．一度目の実験では，被験者が何らかの興味に基づき情報収集し，興味ベクトルの抽出を行う．情報収集の対象となる情報は，150 個の Web サイトを用いる．Web サイトへの属性ベクトルの設定は被験者でない複数の学生が行う．使用するジャンルは「野球」,「コンピュータ」など 20 種類であり，その 20 種類のジャンルに計 10 ポイントを振り分けて各 Web サイトの属性ベクトルを設定する．被験者は 14 名の男性の学生である．情報収集に用いる興味を 3 つ以上に限定し，被験者はその興味に基づき 50 個前後の Web サイトを選択する．Web サイトの選択後，検証のために実装したクラスタリング用アプリケーションを用いて興

表 1: 属性ベクトルの最近ベクトルとの距離

| | 全属性ベクトル | | | 各クラスタ | |
|---|---|---|---|---|---|
| | 平均 | 最大距離 | 最小距離 | 最大距離の平均 | 最小距離の平均 |
| 一度目の実験 | 2.498253 | 12.961481 | 0 | 5.431623 | 0.919788 |
| 二度目の実験 | 2.660437 | 12.569805 | 0 | 6.033251 | 1.032873 |
| 全体 | 2.576642 | 12.961481 | 0 | 5.722410 | 0.974446 |

表 2: 興味ベクトル間のコサイン類似度と距離

| | コサイン類似度 | | | | 重心間距離 | | | |
|---|---|---|---|---|---|---|---|---|
| | 最大値 | 最小値 | 平均値 | 中央値 | 最大値 | 最小値 | 平均値 | 中央値 |
| 全体 | 0.998146 | 0.003534 | 0.866735 | 0.936719 | 10.983624 | 0.562114 | 2.961686 | 2.555631 |
| $\cos\frac{\pi}{6}$ 以上 | 0.998146 | 0.866154 | 0.946154 | 0.956687 | 4.158317 | 0.562114 | 2.143337 | 2.291042 |

味ベクトルを抽出した後，アンケートを行う．アンケートでは，いくつの興味を基に情報を収集したか，抽出された興味ベクトルの値からどのような興味であると連想できるか，という二点を調査する．また，クラスタリングされた属性ベクトル間の距離をもとに，被験者および興味ごとにクラスタの広がりがどれだけ異なるかということを検証する．

二度目の実験は，一度目の実験と同じ内容で 3ヶ月後に行われた．ただし，アンケートよる調査は行っていない．一度目の実験で得られる興味ベクトルと二度目の実験で得られる興味ベクトルを比較し，同じ興味に基づく興味ベクトルを抽出する．そしてそれぞれの興味ベクトル間の近似を検証する．

よって，今回の実験は「アンケートによる，情報整理における興味ベクトルの有用性の検証」,「クラスタの広がりをもとにした手動クラスタリングの妥当性の検証」そして「期間を隔てた興味ベクトル間の同一性の検証」の三つに大別できる．

### 4.2 アンケート結果

一度目の実験の結果，61 個の興味ベクトルが抽出された．そのうち，抽出された興味ベクトルの値から興味を連想できるかという問いに対して，被験者が連想できると答えたのは 50 個となり，全体の 81.96 % となった．このことから，本手法により抽出された興味ベクトルは被験者の持つ興味のイメージをほぼ数値化できると考えられる．

また，アンケートの結果より「情報収集前に思い浮かべた興味の数」と「クラスタリングの結果抽出された興味の数」が異なったのは 14 名中 1 名であった．このことから，稀ではあるが潜在的な興味を抽出できることがわかる．

### 4.3 クラスタの広がり

本節では，クラスタリングされた属性ベクトル間の距離について考察する．クラスタ $C$，属性ベクトル $p, q \in C$ があるとき，属性ベクトル $p, q$ 間の距離を $d_{pq}$ と表す．すると，属性ベクトル $p$ にとって，もっとも近い属性ベクトルとの距離 $d_p^{\min}$ は以下のように定義できる．

$$d_p^{\min} = \min_{p,q \in C} d_{pq}$$

本考察では，まず実験で抽出された全クラスタ内の各属性ベクトルにおける $d_p^{\min}$ を算出する．以降，$d_p^{\min}$ を属性ベクトル $p$ の最近ベクトルとの距離とよぶ．本考察では，全属性ベクトルにおける，最近ベクトルとの距離の平均，最大，最小および，各クラスタにおける最大の最近ベクトルとの距離の平均，最小の最近ベクトルとの距離の平均について考察する．最大の最近ベクトルとの距離の平均や全体の平均の差などから，クラスタの広がりがクラスタごとにどれだけ異なるかを考察する．

前述の考察項目について計算した結果が表 1 である．平均を尺度として結果の考察を行うと，最近ベクトルとの距離の最大は平均距離の 4 倍以上あり，最近ベクトルとの距離の最小は 0 となっている．また各クラスタにおける最大の最近ベクトルとの距離の平均は平均距離の約 2 倍，最小の最近ベクトルとの距離の平均は平均の約 1/2 倍となった．この結果から，クラスタによってその広がり具合に大きな差があることがわかる．よって，ユーザの興味に基づきクラスタリングを行うとき，一定の閾値を用いる階層的クラスタリング手法を適用することは不適切であり，提案した手法が有用であることがわかる．

### 4.4 期間を隔てて抽出された興味ベクトルの同一性

本節では，一度目の実験と二度目の実験により抽出された興味ベクトル間の同一性について検証する．同一性を検証するための尺度として，一度目の実験で抽出された興味ベクトルと二度目の実験で抽出された興味ベクトルとの間のコサイン類似度および距離 (重心間距離) を用いる．コサイン類似度は 1 に近いほど，重心間距離は 0 に近いほど，期間を隔てての興味ベクトルの同一性が高いことになる．また，今回の実験では負の属性ベクトルを考慮していないため，コサイン近似度が 0 に近いほど，重心間距離が大きいほど，同一性が低いことになる．

一度目の実験で抽出された興味ベクトルは 61 個，二度目の実験で抽出された興味ベクトルは 58 個ある．二度目の実験で抽出された興味ベクトルの中で，一度目の実験と同じ興味に基づいたと考えられる興味ベクトルは，49 個であった．同じと考えられる興味ベクトルには，名前付けにおいて，一度目の実験で「スポーツ」とし，二度目の実験で「野球」というように興味が特化しているものも含む．

コサイン類似度および重心間距離の最大値，最小値，平均値，中央値を表 2 に示す．コサイン類似度の最小値は 0 に近く，重心間距離の最大値も，最小値に比べると大変大きなものになっている．ここで，興味ベクトル間の同一性を判断するために，コサイン類似度の閾値に $\cos\frac{\pi}{6}(\simeq 0.866025)$ を用いて検証する．コサイン類似度がコサイン類似度が $\cos\frac{\pi}{6}$ より大きいとき，興味ベクトル間の同一性が高いと判断する．$\cos\frac{\pi}{6}$ より大きい興味

ベクトルの組は 49 組中 36 組あり，これは全体の 73.47 % を占める．また，抽出された興味ベクトルの組全体のコサイン類似度および重心間距離の中央値と，コサイン類似度が $\cos\frac{\pi}{6}$ より大きい興味ベクトルの組全体のコサイン類似度および重心間距離の平均値との差はそれぞれ小さい．このことから全体としてコサイン類似度は 1 に，重心間距離は 0 に近いことがわかる．

これらの結果より，3ヶ月経ってもほとんどの興味ベクトルの変化が小さいことがわかる．よって，興味ベクトルを用いた情報整理をすると，3ヶ月後でも検索にほぼ支障を来さないことが考えられる．

### 4.5 実験やその結果からわかった本手法の問題

情報収集者にとって，収集した情報の属性ベクトルが最適であるという保証がない．これは情報作成者の主観によって属性ベクトルが設定されているためである．そのため，情報収集者と情報作成者ではその情報に対するイメージにギャップが発生する．情報収集者が手を加えることなく情報の属性ベクトルを情報収集者に最適なものに設定することは困難であり，現段階で明確な解決策は無い．ただし，複数人による属性ベクトルの設定から平均を求め，それをその情報の属性ベクトルとすることで，妥当性は保証できると考えられる．

今回の実験ではジャンルを 20 種類に設定したが，これが十分な粒度であったかについて問題がある．粒度が荒いと「スポーツ」であるものが，細かくなると「野球」，「バスケットボール」，さらに細かくなると,「セ・リーグ」,「MLB」のようになる．この点については，全ジャンルの粒度を一定にする必要がある．属性ベクトルの設定方法にも影響するので，今後さらに検討してゆく．

主成分分析法を適用して多次元空間上の属性ベクトルを二次元平面上で表現することで，複数のクラスタが重なることがある．これは囲みにくさや，属性ベクトルの塊がクラスタであるということの認識に影響し，本来の興味ベクトルの値が得にくくなる．囲みやすさを維持するために，三次元に次元を増すことで，少しでも位置関係の欠損を減らし，クラスタの重なりを解消することが解決策として考えられる．

## 5. 既存研究

本研究と類似した既存研究として,「俺デスク」 [4] を挙げる．俺デスクは，ファイル編集や Web サイト閲覧などといった，ユーザのパーソナルコンピュータの操作履歴を管理し，データ着目度やデータ関連度などを用いて情報の検索を容易にするツールである．また，ユーザに特別な手動操作を要求せず，履歴情報の想起を効率化することを目的として研究している．対象となるのは，ローカルディスクに保存された文書や画像，映像などのデータおよび Web サイトなどの Web 上のデータである．データ間の関連は，データを参照した時間の間隔などから関連度を算出する．また，ユーザとデータの関連は，データの参照頻度などからデータ着目度として算出する．俺デスクでは情報の分類分けを行うわけではないので，情報の整理が行えるということではない．また，パーソナルコンピュータで扱えるデータにしか適用できていないという問題点がある．

一方，本論文で提案する手法は，情報の分類分けを行うことで情報整理をする．扱うデータには属性ベクトルが設定されているため，データ間の距離からデータ間の関連度を計ることができる．また，興味ベクトルと属性ベクトル間の距離を基に，どのような興味に基づいて情報を収集したかがわかるので，ユーザとデータとの関連が生まれる．さらに，属性ベクトルさえ設定できればパーソナルコンピュータ外の情報，たとえば電車の吊り広告などに関しても情報の整理が行える．システムへの入力方法としては，RFID や QR コードなどが考えられる．しかし，本手法ではユーザ自身がクラスタリングを行わなければならず，また情報に対して何らかの方法で情報の属性ベクトルを設定しなければならないという手間がかかることが問題点として挙げられる．

## 6. おわりに

本論文では，人の収集した情報の履歴から人の興味を多次元空間上のベクトルとして定量的に抽出する手法を提案した．興味を情報整理のための基準とすることで，その人にとって適切な情報整理が可能になる．

興味のベクトルの有用性，手動によるクラスタリングの妥当性の検証を行った．その結果，実施したアンケートにおいて，被験者は抽出されたすべての興味のベクトルのうち，81.96 % の興味のベクトルに対してその興味を連想できると回答した．抽出された興味のベクトルは，収集された情報の検索において 3ヶ月後も有用であることがわかった．また，得られたクラスタから，手動によるクラスタリングは情報を人の興味ごとにクラスタリングすることに適していることがわかった．しかし，多次元空間の軸として使用するジャンルや情報の属性ベクトルの妥当性に問題がある．実験で使用したジャンルは粒度が一定ではなく，網羅的でないので属性ベクトルを設定する自由度が制限される結果になった．情報の属性ベクトルは，情報作成者が主観により設定するため，情報に対する印象が情報収集者と異なる場合がある．今後は，網羅的なジャンルを一定の粒度で扱い，客観的な属性ベクトルを設定する手法の研究を行う予定である．

## 参考文献

[1] 総務省 情報通信政策局 情報通信経済室，情報通信による経済成長に関する調査 報告書，http://www.johotsusintokei.soumu.go.jp/linkdata/other014_200707_hokoku.pdf，2008 年 1 月 28 日閲覧

[2] 齋藤 堯幸，宿久 洋，関連性データの解析法 多次元尺度構成法とクラスター分析法，p.134-136，p180-182，共立出版 (2006).

[3] 田中 豊，脇本 和昌，多変量統計解析法，現代数学社，p.53-99 (1983).

[4] 大澤 亮，高汐 一紀，徳田 英幸，俺デスク:ユーザ操作履歴に基づく情報想起支援ツール，情報処理学会 第 47 回プログラミング・シンポジウム (2006).