D-040

多次元ベクトル表現を用いた興味のクラスタリング
Clustering of User Interests with Multidimentional Vector Representation

樋口 賢治† 江坂 直紀‡ 原田 史子† 島川 博光†
Kenji Higuchi Naoki Esaka Fumiko Harada Hiromitsu Simakawa

1. はじめに

今日，インターネット回線の高速化，記憶媒体の大容量化が進んでいる．実世界においても，町を歩けばビルの壁や駅のホームなど，あらゆる所に広告が存在し，個人の触れる情報の量が増えている．我々の生活はさまざまな種類，膨大な量の情報に溢れ，収集した情報の整理や活用は困難である．

これは，収集した情報を，適切なカテゴリに分けることができないためである．収集した情報を適切なカテゴリに分けることができると，情報整理にかかるユーザの負担を軽減できる．あわせて，情報整理は，情報の利活用を促進する．

我々は，収集した情報のカテゴリ分けを行うにあたり，その情報を収集したときのユーザの興味を使用する．ユーザが収集する情報は，その情報がさまざまな分野の中で，どの分野に，どの程度関連しているかを示すベクトルで表されると想定する．本論文では，ユーザが情報を収集した履歴から，興味を多次元空間上のベクトルとして同定する手法を提案する．

2. 情報の整理

2.1 情報のカテゴリ分け

人間はある興味を持ち，その興味に基づいて情報を収集する．人間の興味は複数あり，収集した情報の中には同じ興味を基に収集したものがあれば，違う興味を基に収集したものもある．すなわち，収集した情報は収集したときの興味別に分けることができる．

そこで，収集した情報をカテゴリに分ける尺度として，ユーザが情報を集めるときに持っている興味を使用する．そうすることで，その情報を収集した目的がユーザにとってわかりやすくなるため，収集した情報の活用が容易になる．本研究は，この興味を過去収集した情報の履歴から定量的に求めることを目的とする．

2.2 興味ベクトル

我々は，ユーザの各興味が，どの分野にどの程度関連しているかを示すベクトルが存在すると仮定する．たとえば，ユーザの興味を示すベクトルの分野に「スポーツ」，「ファッション」，「ショッピング」があるとする．このとき，ユーザのある興味が「スポーツ」に 0.7，「ファッション」に 0.3，「ショッピング」に 0.0 関連しているとすると，そのベクトルは

$$\begin{bmatrix} \text{スポーツ} \\ \text{ファッション} \\ \text{ショッピング} \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.3 \\ 0.0 \end{bmatrix}$$

となる．このように，すべての情報が，各分野にどれだけ関連するかを示すベクトルが設定されていると想定する．この分野を以降，属性とよぶことにする．ユーザがある興味を持って情報を収集したとすると，ユーザが収集する情報のベクトルはその興味のベクトルの近くに集まると考えられる．いいかえれば，ユーザの興味のベクトルは収集した情報のベクトルの集中したところに存在することがわかる．よって，ユーザの収集した情報のベクトルの履歴からベクトルが集中して存在する領域を見つけることができれば，ユーザの興味のベクトルが求められる．

情報のベクトルは定量的に表されるため，そこから求められた興味のベクトルも定量的に表すことができる．このように，定量化された興味のベクトル表現を，興味ベクトル [1] とよぶことにする．

2.3 既存研究

類似する既存研究として，Behavior Targeting[2]，俺デスク [3] の 2 つがある．

Behavior Targeting はユーザの行動履歴から広告の配信を最適化する広告手法である．ウェブサイト訪問履歴を使用し，ユーザが興味・関心をもつ特定分野を同定し，その分野に関係する広告を個人向けに配信する．Behavior Targeting はユーザのウェブサイト訪問履歴を利用し，興味を導いている．しかし，広く一般の情報にこの手法を適用することはできない．

俺デスクはユーザの操作履歴からユーザのデータに対する着目度の度合いやデータの関連度を示し，情報の検索を容易にするツールである．履歴から情報の検索を容易にするが，俺デスクはユーザの興味を求めているというわけではない．興味を自動的に推定し，興味に基づいて情報をカテゴリ分けできなければ，ユーザの情報整理の負担は軽減できない．

3. 興味ベクトルの同定

3.1 手法の概要

本研究が提案する，ユーザの収集した情報の履歴から興味ベクトルを同定する手順は，以下のようになる．

まず，システムはユーザの集めたすべての情報の属性とその値をもとに，情報を多次元空間に配置する．次に，それらの情報に対し主成分分析法 [4] を適用し，多次元表現を二次元表現で表し可視化する．最後に，平面上に描画された点をユーザの手によりクラスタリングする．システムはそのクラスタの重心を同定し，興味ベクトルとして導く．この一連の流れを図 1 に示す．

本手法は，情報の持つベクトルからユーザの興味を同定できる．すなわち，ベクトルが定義されていれば，パソコン上の音楽や動画，画像データだけでなく，電車の吊り広告やポスターなど，実世界に存在するものにも適用できる．

†立命館大学 情報理工学部 情報システム学科
‡立命館大学大学院 理工学研究科

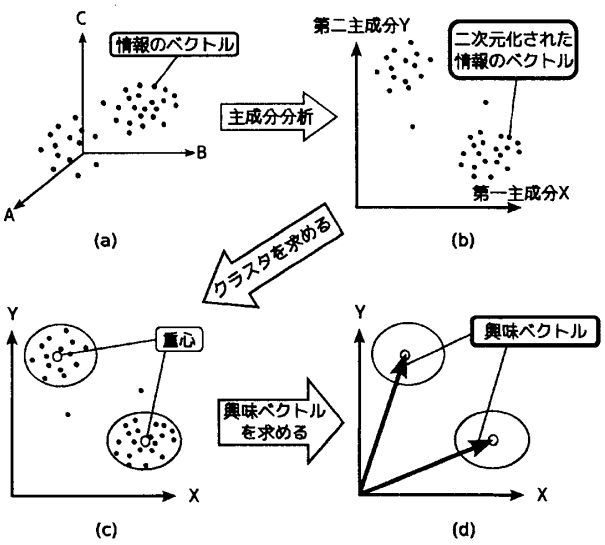

図 1: 興味ベクトルの同定

3.2 情報の収集

ユーザの収集するすべての情報の属性とその値は情報作成者の主観により設定される．値には上限が定められているとする．

ユーザは自分の興味がある情報を収集し，その情報は時系列として管理される．収集された情報は，属性値をもとに，全属性を軸としてもつ多次元空間上に配置される．図 1 (a) の例では属性として A，B，C の三次元を挙げているが，実際は三次元より大きな次元で考える．

3.3 主成分分析法の適用

システムはユーザがクラスタリングを行えるように，情報を配置した多次元空間を二次元平面で表し，描画する．多次元表現を二次元表現で表すために，収集した情報のベクトルに主成分分析法を適用し，第一主成分および第二主成分を求める．各情報のベクトルを第一主成分，第二主成分を元に変換し，この 2 つの主成分を軸として二次元平面として描画する．図 1 (b) の例では主成分の軸を X，Y として挙げている．

寄与率の高いこの 2 つの主成分をもとに多次元空間を二次元で表現することで，ベクトルのもとの位置関係をほとんど維持できる．

3.4 クラスタ分け

二次元平面に描画された情報のベクトルについて，ユーザは主観により同じクラスタに属すると考えられる情報のベクトル群をフリーハンドで囲い，クラスタ分けする．図 1 (c) に示すように，システムはその囲われた情報をクラスタとして把握し，その重心を求める．

ここで，このクラスタというのはユーザの興味がある情報の集合なので，そのクラスタの重心はユーザの興味の集中しているベクトルと考えられる．よって各クラスタの重心がユーザの興味ベクトルであるとみなす．

クラスタは複数個存在すると考えられるため，情報のベクトルを二次元化することでクラスタが隠れることが想定される．しかし，情報を収集する期間を短い期間に区切ることで，各期間におけるユーザの興味は多くても 2 つ程度に収まる可能性がある．つまり，期間を区切って処理を行うことでこの問題は回避できる．その場合，ある期間で得られたクラスタの中に，以前抽出したクラスタに近いものがあるとすれば，そのクラスタとの合成を行わなければならない．

4. 展示会での適用例

本手法は，前述した通り，あらゆる情報に対して適用できる．ここで，1 つの例を示す．

あるユーザが World PC Expo などの，各出展者がブースに別れて展示するスタイルの展示会において，訪問先のいろいろなブースにてパンフレットを貰い，結果として過度にパンフレットが集まったということを想定する．パンフレットは多数あり，パンフレットの分類が困難であるため，ユーザは整理できないでいる．そこでユーザはパンフレットを分類するため，本手法を適用する．

ブース出展者が独自に自分のブースの属する分野とその値を設定しており，それがパンフレット上に印刷された QR コードから読み取れるものとする．値は合計 10 ポイントで振り分けられている．ユーザは家に帰りパンフレットの QR コードをシステムに読み込ませる．するとシステムは端末の画面上にパンフレットのベクトルを表示し，そのベクトルをユーザがクラスタに分ける．システムはクラスタの重心を求め，それを興味ベクトルとしてユーザに提示する．

興味ベクトルを求めることで，パンフレットのベクトルと興味ベクトルの距離から情報を自動でカテゴリ分けできるようになる．結果としてユーザはパンフレットを自分の興味に対応して整理できるようになる．

5. おわりに

本論文では，ユーザが情報を収集した履歴から，ユーザの興味を多次元空間上のベクトルとして同定する手法を提案した．本手法では，収集した情報のクラスタリングをユーザ自身が行わなければならないという問題がある．自動的に最適なクラスタリングを行える手法を適用することが今後の課題である．

参考文献

[1] 江坂 直紀，高田 秀志，島川 博光，Pen-less Recorder による情報収集に基づく興味ベクトルの把握，組込みシステムシンポジウム 2006，pp.132-135，2006

[2] Bill Harvey，Next Century Media，The Advantage of Behavior Targeting Increases Dramatically with Frequency，http://www.tacoda.com/assets/TACODA_Advantage_of_Behavior_Targeting.pdf，2006

[3] 大澤 亮，高汐 一紀，徳田 英幸，俺デスク:ユーザ操作に基づく参照履歴検索ソフトウェア，情報処理学会第 47 回プログラミング・シンポジウム，pp.219-220，2006

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