

ユーザが抱く印象を用いた個人嗜好の表現 —Folksonomyを利用したユーザの印象の推測方法—

田上 道士¹ 山場 久昭¹ 高塚 佳代子¹ 岡崎 直宣¹ 富田 重幸¹

概要: 本研究では、「ユーザがアイテムに対して抱く印象」を用いた人の好みの表現を提案する。アイテムに対して実際にユーザがどのような印象を抱いたかはわからないので、Folksonomyにおけるタグの情報をを用いて、印象の推測を行う。この推測した印象を「コンセプト」と呼ぶ。以前の研究で提案した方法で実際に導出されたコンセプトの妥当性を調べた上で、新たに改良したコンセプトの導出方法を提案する。そして、この新たな導出方法の妥当性と性能の向上を調べるために旧手法との比較実験を行い、また、人の好みの表現にコンセプトを用いることの有用性を調べるために推薦実験を行った。

Representation of human preference using users' impressions —A method of inference of users' impressions using folksonomy—

TANOUE MICHIHITO¹ YAMABA HISAAKI¹ TAKATSUKA KAYOKO¹ OKAZAKI NAONOBU¹
TOMITA SHIGEYUKI¹

Abstract: This paper proposes a method to represent human preference using impressions of items derived from users. Since such impressions are not known explicitly, we infer the impressions from information of tagging in a folksonomy. An inferred impression is called a "concept." In this study, we investigate validity of the concepts obtained by our previous method, and propose the new improved derivation method of concepts based on the results of the investigation. Experiments were carried out in order to evaluate the performance of the improved method through comparison with the previous method, and a recommendation experiment was carried out in order to confirm the usefulness of the representation of human preference using the concepts obtained by the new method.

1. 緒言

近年、ユーザにとって有用な情報を自動的に提示する「推薦システム」の重要性が増しており、このシステムが適切な情報推薦を達成するためには、人の好みの適切な表現（または把握）が非常に重要である。従来の推薦手法の多くでは、推薦者側があらかじめ定義した属性やカテゴリが用いて人の好みが表示されていたが、多数のユーザのアイテムに対する感覚や認識、印象を反映させることにより、より適切な人の好みの表現が期待できる。

以前の研究 [1][2] で、筆者らは Serendipity を考慮した推薦の実現を目指し、「コンセプト」と呼ぶ考えを導入して

いる。この研究では、ユーザ達がアイテムに対して抱く印象を推測し、その推測結果を用いた推薦手法を提案した。この推測した印象が「コンセプト」である。この推測を行うにあたって、Folksonomyにおけるタグの情報をを用いている。

本研究では、コンセプトを用いて人の好みを表示することを目指す。「ユーザ達がアイテムに対して抱く印象」を推測したものであるコンセプトを用いて表現した人の好みは、「あらかじめ定義された属性」を用いる場合よりも、より適切であると期待できる。

ただし、このような人の好みの表現を実現するためには、コンセプトが妥当なものである必要がある。そこで本研究では、まず、以前提案した手法で得られたコンセプトの妥当性を調べた。その結果、ほぼ期待するようなコンセプト

¹ 宮崎大学
University of Miyazaki

が得られていることを確認できたが、一方で、いくつかのコンセプトは人の好みを表現するのに不十分な点があった。

そこで、コンセプトの精度を向上させるため、コンセプトの形成方法の改良を行った。そして、改良したコンセプト形成方法の性能を評価するために、以前の手法との比較実験を行った。さらに、改良手法で導出されたコンセプトを用いて表現した人の好みの妥当性を調べるための推薦実験を行った。

2. 人の好みの表現

本研究では、「ユーザがアイテムに対して抱く印象」を用いた人の好みの表現を提案する。具体的には、以前の研究で Folksonomy におけるタグの情報を利用して導入した「コンセプト」と呼ぶ考えを活用する。

2.1 コンセプトと呼ぶ考え

従来の推薦手法の多くは、人の好みを表現するにあたって、推薦者側（例えば、推薦対象となるアイテムに関しての専門家）があらかじめ定義した属性やカテゴリを用いていた。従来の推薦手法は、協調フィルタリングと内容ベースフィルタリングに大別され [3][4][5]、特に後者では、アイテムの内容・特徴に基づいて推薦されるアイテムが選ばれるが、その際に推薦者側があらかじめ定義した属性（例えば、著者やジャンルなど）を用いてアイテムの特徴は表現されることが多く、ユーザの好みもまた、それと同じ属性で表現されている。

しかし、人の好みの表現には、「あらかじめ定義された属性」を用いるよりも、「ユーザ達のアイテムに対する印象」を用いる方がより好ましいといえる。なぜなら、ユーザ達自身が感じている多様な視点が反映されることによって、人の好みのより適切な表現が期待できるためである。

以前の研究で、筆者は Serendipity を考慮した推薦の実現を目指し、ユーザ達がアイテムに対して抱く印象を用いて、アイテムの特徴の表現した。そして、「推薦を受けるユーザが好きなアイテムに対してユーザ達が抱く印象と、同じ印象が抱かれているアイテム」を推薦する手法を提案した。ここでの「印象」とは、アイテム（の内容や特徴、テーマ、トピックなど）に対してユーザが感じるもの一つ一つを示す。例えば、ある本に対して「ミステリー小説だ」「感動するストーリーだ」などが抱かれているとき、それら一つ一つのことを指す。

ただし、実際にユーザがどのような印象を抱いたかはわからないので、Folksonomy におけるタグの情報を用いて、ユーザが抱いた印象の推測を行った。この推測された印象を「コンセプト」と呼ぶ。このコンセプトを用いてアイテムの特徴、及び、ユーザの好みを表現した。

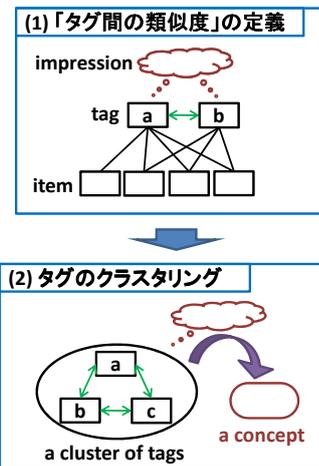


図 1 コンセプトの形成

Fig. 1 Derivation of concepts.

2.2 コンセプトの形成

コンセプトの形成は次のように行う（図 1）。「ユーザはアイテムに対して抱く印象に基づいて使用するタグを選ぶ」ものと仮定する。

- (1) まず、「タグ間の類似度」を定義する。先程の仮定から、ある二つのタグが同じ印象に基づいて使用されているとき、それらタグは共起しやすい（同じアイテムに付与されやすい）と考えられるので、タグ間の共起情報を用いて、この類似度を計算する。
- (2) 互いに高い類似度を持つタグ同士が集まったクラスタ群を生成する。このようなクラスタ内のタグ群を、共通の印象に基づいて使用されているだろうタグとみなし、できたクラスタそれぞれを一つのコンセプトと対応付ける。

2.3 コンセプトを用いた人の好みの表現

あるユーザの好みは「得られた各コンセプトへの関心の強さ」で表現する。ユーザのコンセプトへの関心強さは、「そのユーザの好きなアイテムに付けられているタグと、そのコンセプトとの間の関連強さ」に基づいて、算出する（図 2）。具体的には、まず、「タグとコンセプトとの間の関連強さ」を、そのタグと、そのコンセプトに対応するクラスタ内の各タグとの類似度に基づいて、算出する。あるアイテムに付けられているタグがあるコンセプトと関連強いとき、そのコンセプトに対応する印象に基づいてユーザ達はそのタグを付与したと考えられる。このため、次に、「アイテムとコンセプトとの間の関連強さ」を、そのアイテムに付けられているタグと、そのコンセプトとの間の関連強さに基づいて算出する。そして、「ユーザのコンセプトへの関心強さ」を、そのユーザが好きなアイテムと、そのコンセプトとの関連強さに基づいて算出する。

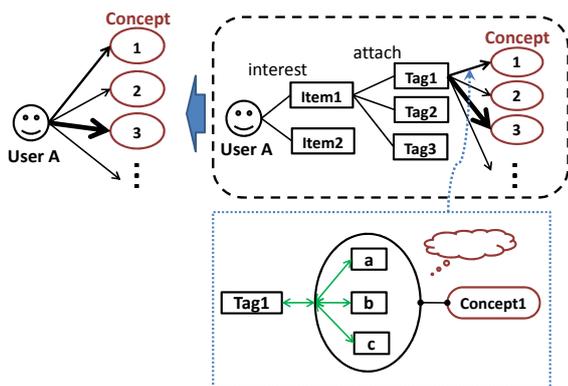


図 2 コンセプトを用いた人の好みの表現

Fig. 2 Representation of human preference using concepts.

3. 以前の研究で提案した手法で得られたコンセプトの妥当性の分析

前述したような人の好みの表現方法を実現するためには、コンセプトが妥当なものである必要がある。そこで、以前の研究で提案した手法を用いて実際に得られたコンセプトが妥当なものであるかどうかを調べる。具体的には、各コンセプトに対応したクラスタ内のタグ群から、ユーザが感じたであろう印象が連想できれば、妥当なコンセプトが得られていると判断することとする。逆に言えば、クラスタ内のタグ間に関連性が見出せず、ユーザが抱いているであろう印象の連想が困難であるときは、そのコンセプトは妥当でないと判断することとした。また、ユーザの好みの表現に用いる「タグとコンセプトとの間の関連度」にも焦点を当てる。あるタグがあるコンセプトに関連強いとき、その関連強さが高いほど、そのコンセプトに対応する印象に基づいてそのタグは使用されていると期待できるので、各コンセプトと関連強いタグ群から、ユーザが抱いたであろう印象を連想できれば、妥当なコンセプトが得られていると判断することとする。さらに、本研究の基本的な考え方下では、ユーザが抱いた印象一つに対して一つのコンセプトが対応するので、同じ印象に対応するコンセプトが複数得られることは妥当でないと考えられる。このため、そのようなことが起きてしまっていないかどうかを調べる。

3.1 方法

この分析に用いるデータは「ブックログ (Booklog)」 (<http://booklog.jp/>) から収集した。ブックログとは仮想本棚を作成できる Web サービスである。得られた本数は 6,717 冊、その本に付けられていたタグは 18,922 個だった。今回、コンセプトへのタグの影響を見やすくするために、7 冊以上の本に付けられているタグに絞った (1,115 個)。

このデータから、以前の研究で提案した手法を用いて、コンセプトの形成を行った。そして、以下の (i) と (ii) の

表 1 各クラスタ内のタグ群の例

Table 1 Examples of tags in each cluster.

Cluster	Part of tags
1	IT, プログラミング, 技術書, オライリー, C, C 言語
2	IT, プログラミング, 技術書, オライリー, Java, java
3	it, プログラミング, 技術書, JavaScript, javascript, Javascript
4	IT, プログラミング, 技術書, JavaScript, javascript, Javascript

表 2 各コンセプトと関連強い上位タグ群の例

Table 2 Examples of top tags that have high relevance with each concept.

Concept	Top tags
1	C, C 言語, プログラミング, Ruby
2	Java, java, programming, Android
3	Javascript, JavaScript, javascript, Java
4	Javascript, javascript, JavaScript, Java

タグ群それぞれを見て、ユーザが抱いたであろう印象を連想できるかの判断によって、そのコンセプトの妥当性を調べた。

- (i) 各コンセプトに対応したクラスタ内のタグ群
- (ii) 各コンセプトと関連強い上位タグ群

また、同じ印象に対応するコンセプトが複数得られていないかを調べるために、互いに似たタグ構成のものが複数できていないかを調べた。

3.2 結果

(i) の例を表 1 に、(ii) の例を表 2 に示す。表 2 では、関連度が高いタグほど左の方に表示している。

被験者は、(i) と (ii) のどちらにおいても、ほとんどのコンセプトが妥当なものであると判定した。例えば、表 1 の Cluster1 と表 2 の Concept1 からは共に、「C 言語」という印象を抱いたと判定できた。

一方で、同じ印象に対応するであろうコンセプトが複数できてしまっていた。例えば、表 2 の Concept3 と 4 のように、同じタグ群と関連強いコンセプトがいくつか存在した。また、表 1 の Cluster1 と 2, 3 と 4 のように、その大半が同じタグ群で構成されるクラスタがいくつか存在した。

3.3 考察

今回の分析の中で、二つのクラスタが「互いに似ている」(図 3 に示すように、それら二つのクラスタが多くのタグを共有している) とき、その二つのクラスタの中から取り出した二つのタグの間の類似について、以下の知見が得られた。すなわち、互いに似ている二つのクラスタを選んで、そのクラスタ内のいずれかのタグを二つ選び、その類似度を算出した結果から、以下二つに該当するタグの組は、それらに該当しないタグの組よりも、類似度が比較的高い傾

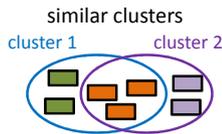


図 3 互いに似ているクラスタ
 Fig. 3 Similar clusters.

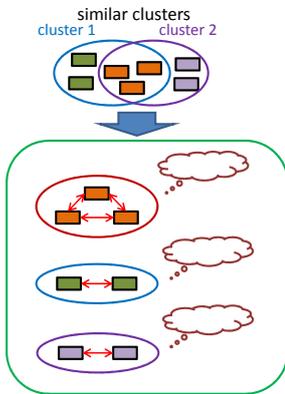


図 4 目標とするクラスタ
 Fig. 4 Target of clusters.

向にあることがわかった。

- (α) 一方のクラスタにしか含まれていないタグの組
 (図 3 の緑のタグの組, 或いは, 紫のタグの組)
- (β) それぞれのクラスタに共に含まれるタグの組
 (図 3 のオレンジのタグの組)

例えば, 表 1 の Cluster1 と 2 に注目したとき, Cluster1 にしか含まれていないタグの組「C, C 言語」や Cluster2 にしか含まれていないタグの組「Java, java」それぞれの類似度は, (α)(β) に該当しないタグの組 (例えば「プログラミング, C」など) の類似度よりも高い。

このことから, “互いに特に高い類似度を持つようなタグ同士からなるクラスタ群” が生成できれば, 「(α) に該当するタグ群」と「(β) に該当するタグ群」とを分けることで, 互いに似ているクラスタの生成を避けられると期待できる (図 4)。ただし, 「特に高い類似度」の高さの程度は各タグによって異なるため, 工夫が必要である。

4. コンセプトの形成方法の改良

コンセプトの精度向上のために改良した, 新たなコンセプトの形成方法を提案する。具体的には, タグ間の類似度の定義と, タグのクラスタリング手法, 及び, タグとコンセプトとの間の関連度の計算方法を改良した。

4.1 タグ間の類似度

まず, タグ間の類似度は, 「ある二つのタグについて, 一方のタグが共起しやすいタグ群に, 他方のタグも共起しやすいとき, それらタグは共通の印象に基づいて使用されている」と仮定し, 各タグの共起パターン間の類似度に基づ

いて定義する。以前の研究では, 二つのタグが共起しやすいかどうか (以下, この尺度を「タグ間の共起度」と呼ぶ) に基づいて, タグ間の類似度を定義した。しかしこの定義では, タグ間の類似度の高さがそれらタグの発生頻度に依存してしまっていた。例えば, 同じ意味で用いられているタグ同士であっても, それら各タグの発生頻度が低ければ, そのタグ間の類似度も低くなってしまふ。

各タグの共起パターン (各タグの, 全てのタグそれぞれとの共起度からなるパターン) を, タグ間の類似度の算出に適用することで, 単にタグ間の共起度を用いた場合よりも, 同じ印象に基づいて使用されているタグ同士の類似度がより高くなることが期待できる。また, 各タグの発生頻度による影響を避けることが期待できる。

具体的には, まず, 二つのタグ間の共起度を導入するため, 各アイテムに対する二つのタグ (a, b) 間の付与関係を以下の四つにまとめる。

- (A, B) 両方のタグが付けられている
 - (\bar{A}, \bar{B}) どちらのタグも付けられていない
 - (A, \bar{B}) a は付けられているが, b は付けられていない
 - (\bar{A}, B) b は付けられているが, a は付けられていない
- 「(A, B) と (\bar{A}, \bar{B}) が起こりやすく, (A, \bar{B}) と (\bar{A}, B) が起こりにくい」なら, a, b は共起しやすいといえる。このため本研究では, タグ間の共起度として AEMI (Augmented Expected Mutual Information) [6] を採用する。

$$AEMI(a, b) = MI(A, B) + MI(\bar{A}, \bar{B}) - MI(A, \bar{B}) - MI(\bar{A}, B)$$

ここで, A は「アイテムにタグ a が付けられている」という事象を, B は「アイテムにタグ b が付けられている」という事象を示す。 \bar{A} は「アイテムにタグ a が付けられていない」という事象を示す。また, $MI(A, B)$ は「事象 A と B が相互にどのくらい依存しているかの情報量」であり, 以下の式で表される。

$$MI(A, B) = P(A, B) \log \frac{P(A, B)}{P(A)P(B)}$$

$P(A)$ は「事象 A の発生頻度 (タグ a が付けられているアイテムの割合)」を, $P(A, B)$ は「事象 A と B の同時発生頻度 (タグ a と b が共に付けられているアイテムの割合)」を示す。

次に, この共起度を用いて各タグの共起パターンを表現する。全てのタグ間の共起度 $AEMI(t_i, t_j)$ に関して, $N \times N$ (N : タグ数) 次元の行列ができるので, その各行ベクトルを, 対応するタグの共起パターンとする。

$$\vec{v}(t_i) = [AEMI(t_i, t_j)]_{j=1, 2, \dots, N}$$

ここで, $AEMI(a, a)$ はタグ a の発生に関するエントロピーである。

$$AEMI(a, a) = -P(A) \log P(A) - P(\bar{A}) - \log P(\bar{A})$$

そして、タグ a と b の共起パターン間の類似度を、 a と b の間の類似度 $sim(a, b)$ とする．本研究ではコサイン類似度を採用する．

$$sim(a, b) = \frac{\vec{v}(a) \cdot \vec{v}(b)}{\|\vec{v}(a)\| \|\vec{v}(b)\|}$$

4.2 タグのクラスタリング

新たなクラスタリング手法も、基本的には、以前の方法と同様、「ある共通の印象に基づいて使用されているであろうタグ群」を見付けるために、互いに高い類似度を持つタグ同士からなるクラスタを生成する．本研究ではいくつかのアイデアを導入したが、本稿では、その内の一つを4.2.1節で説明する．また、提案するクラスタリング手法全体のアルゴリズムを4.2.2節に示す．

4.2.1 多段クラスタリング

互いに似ているクラスタが生成されることを避けるために、3.3節で述べたような、前回の方法で生成された似たクラスタの「一方のクラスタにしか含まれていないタグ群」と「それぞれのクラスタに共に含まれるタグ群」とが別々のクラスタに含まれるようになることを目指す(図4)．このために、「互いに特に高い類似度を持つタグ同士」からなるクラスタを生成したいが、「特に高い類似度」の高さの程度は各タグによって異なる．

そこで、「互いに特に高い類似度を持つタグ同士」からなるクラスタを生成するために、タグ間の類似度の高さに応じた多段的なクラスタリングを導入する．まず、最初のステップでは、「特に高い類似度」として高めの閾値を設定し、特に類似し合うタグ群からなるクラスタ群を生成する．このステップでできた各クラスタ内のタグは、ある共通の印象に基づいて使用されているタグとみなせるので、以降のステップではそれらタグを除いたタグ群を対象にクラスタリングを行う．次のステップでは、先程よりも低い閾値を新たに設定し、残りのタグ群(以前までのステップでできたクラスタには含まれないタグ群)からクラスタ群を生成する．この処理を数回繰り返す(図5)．

4.2.2 アルゴリズム

初めに以下を用意しておく．

- 生成されたクラスタの集合 C_{set} (初期状態では空集合)
- 全てのタグからなるタグの集合 T
- ステップ数 S
- 閾値 V_k ($k = 1, 2, \dots, S$), ($V_k > V_{k+1}$)

タグ間の類似度が閾値 V_k 以上のタグの組 (t_i, t_j) ($t_i, t_j \in T | i \neq j$) からなるリストを生成する．そのリストから昇順にタグの組を選択し、以下の処理を行う．

- (1) タグ t_i と t_j からなるクラスタ cl を生成する．
- (2) 以下の処理を繰り返す．

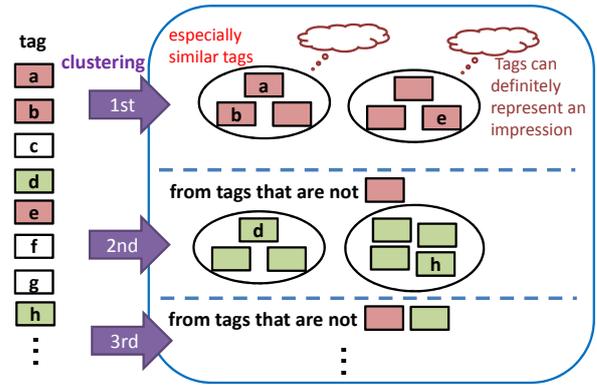


図5 多段的なクラスタリング
Fig. 5 Multi-step clustering.

- cl 内のタグ群に最も類似するタグ t_{max} ($t_{max} \in T, t_{max} \notin cl$) を選ぶ．タグ t とクラスタ cl との間の類似度 $sim(t, cl)$ は、 cl 内の各タグと、 t との間の類似度の平均値とする．
- 以下の閾値 $V(cl)$ を求める．

$$V(cl) = \alpha \times ave_{sim}(cl)$$

ここで、 $ave_{sim}(cl)$ は、 cl に含まれるタグ間の類似度の平均値を示す．また、 α はこの閾値の高さを調節するためのパラメータである．

- $sim(t_{max}, cl) \geq V(cl)$ のときは、 t_{max} を cl に追加し、(i) に戻る．
 - (iii) の条件を満たさなければ、この処理を終了する．
- cl と C_{set} 内の各クラスタ cl_s との間の類似度 $sim(cl, cl_s)$ 、及び、閾値 $V(cl, cl_s)$ を求める． $sim(cl, cl_s)$ は、 cl 内の各タグと cl_s 内の各タグとの間の類似度の平均値とする．また、 $V(cl, cl_s)$ は $V(cl)$ と $V(cl_s)$ の平均値とする．
 - C_{set} 内に、 cl と類似するクラスタ cl_{sim} が含まれているとき ($sim(cl, cl_{sim}) > V(cl, cl_{sim})$)、 cl 内のタグ群 ($\notin cl_{sim}$) を cl_{sim} に追加する．
 - 類似するクラスタがなければ、 cl を C_{set} に追加する．

これら処理によって更新された C_{set} 内のクラスタに含まれるタグ群を、 T から取り除く．これを S ステップ繰り返す．

4.3 タグとコンセプトとの間の関連度

タグとコンセプトとの間の関連度は、そのタグとそのコンセプトに対応するクラスタ(内の各タグ)との間の類似度に基づいて計算する．具体的には、まず、全てのタグとクラスタの組み合わせ (t_i, cl_j) それぞれについて、その間の類似度 $sim(t_i, cl_j)$ を求める． $sim(t_i, cl_j)$ は、 t_i と cl_j 内の各タグとの類似度の平均値とする．そうすると、

$sim(t_i, cl_j)$ を要素とする $N \times M$ (N :タグ数, M :クラスタ数) 次元の行列が表現できる。次に, “あいまいさを持つタグ”(多義的, 或いは, 抽象的な意味で使用されているタグ) による影響を避けるため, 各タグに関するベクトルの正規化を行う。基本的には, こうしてできた行列の各要素を, タグと(そのクラスタに対応する)コンセプトの間の関連度 $rel(t_i, co_j)$ とする。

$$rel(t_i, co_j) = rel(t_i, cl_j) = \frac{sim(t_i, cl_j)}{\sqrt{\sum_k sim(t_i, cl_k)^2}}$$

また, 更に“あいまいさを持つタグ”の影響を避けるために, この関連度を拡張する。具体的には, この行列の各要素を二乗した値を, 該当するタグとコンセプトとの間の関連度として採用する。この二乗した値を採用することにより, “あいまいさを持つタグ”であるほど, そのタグとコンセプトとの関連度がより低くなる。実際には, $sim(t_i, cl_j)$ が負の値を取ることもあるため, タグとコンセプトの間の関連度は以下とする。

$$rel(t_i, co_j) = \begin{cases} rel(t_i, cl_j)^2 & \text{if } sim(t_i, cl_j) \geq 0.0 \\ -rel(t_i, cl_j)^2 & \text{else} \end{cases}$$

5. 実験

改良した手法(以下, 新手法)の有効性を調べるために, 以前の研究で提案した手法(以下, 旧手法)との比較実験を行った(5.1~5.3節)。

実験に用いたデータは3節のものと同じである。

5.1 タグ間の類似度に関する実験と結果

まず, 新手法によるタグ間の類似度の精度向上を調べる実験を行った。具体的には, タグ間の類似度の再定義によって, 旧手法における, タグ間の類似度の高さがタグの発生頻度に依存してしまう問題が改善されていることの確認を行った。

5.1.1 方法

まず, いくつかのタグの組を選び出す。これらタグの組に関して, 旧手法を用いた場合と新手法を用いた場合のそれぞれの類似度を求める。そして, 各タグの発生頻度によらず, 同じ意味で用いられているタグの組の類似度が比較的高い値になっているか否かを調べた。

5.1.2 結果

各手法での定義によるタグ間の類似度の値の例を表3に示す。

この結果から, 新手法は, タグ間の類似度の高さがタグの発生頻度に依存してしまう問題が改善されていることが確認できた。新手法では, 同じ意味で用いられているようなタグの組の類似度は, 最大値と比較して, いずれも高い値となっている。比較的発生頻度が低いタグの組である「C, C 言語」や「javascript, Javascript」間の類似度も比

表 3 タグ間の類似度の結果の例

Table 3 Examples of similarity between tags.

A pair of tag	旧手法	新手法
(最大値)	0.1569	0.9983
ミステリー, ミステリ	0.1396	0.9377
C, C 言語	0.0063	0.8837
Java, java	0.0061	0.7341
javascript, Javascript	0.0083	0.9619
プログラミング, 技術書	0.0159	0.6528
プログラミング, C	0.0049	0.4176
プログラミング, Java	0.0058	0.3786
プログラミング, javascript	0.0048	0.4207

表 4 二つのクラスタに共通するタグの数の平均値

Table 4 Average of number of tags that appear commonly in each pair of clusters.

旧手法	新手法
5.1688	0.0019

較の高い値となっていることから, タグ間の類似度の高さがタグの発生頻度に依存してしまう問題の改善が確認できた。一方で, 旧手法の場合は, 同じ意味で用いられているタグの組であっても, 最大値と比べてその類似度が比較的低いタグの組が多かった。例えば, 表3に示すように旧手法だと, 発生頻度が高いタグの組である「ミステリー, ミステリ」間の類似度は比較的高いが, 発生頻度が低いタグの組である「javascript, Javascript」間の類似度は比較的低い。

5.2 タグクラスタに関する実験と結果

次に, 新手法が目指す, 互いに似ているクラスタの生成の抑制が実現できているかを調べる実験を行った。

5.2.1 方法

まず, 旧手法と新手法それぞれを用いてタグのクラスタリングを行う。

そして, 得られたクラスタ群から, クラスタの全ての組み合わせを選ぶ。このクラスタの組それぞれについて, それら二つのクラスタに共に含まれるタグの数を求める。共に含まれるタグがない場合は, 0として数える。そして, この平均値(「クラスタの組それぞれが共有するタグの数の総和」/「クラスタの組の数」)を求め, 旧手法と新手法それぞれの結果を比較する。

5.2.2 結果

平均値の結果(表4)から, 新手法では互いに似ているクラスタ群の生成を回避できていることが確認できた。旧手法よりも, 新手法の方がクラスタ間で共通するタグの数は少ないことがわかる。

表 5 コンセプト間の類似度の結果

Table 5 Results of similarity between concepts.

	旧手法	新手法
平均値	0.3779	0.0185
最大値	0.9994	0.9903
最小値	0.0	-0.0925

5.3 コンセプトに関する実験と結果

新手法によって、類似した（同じ印象に対応する）コンセプトが複数形成されにくくなっているかを調べる実験を行った。

5.3.1 方法

旧手法と新手法それぞれを用いて、コンセプトを形成し、得られたコンセプト間の類似度を求めた。コンセプト間の類似度の計算には、コサイン類似度を採用した。

5.3.2 結果

得られた結果を以下のようにまとめた。まず、各手法で求めたコンセプト間の類似度の平均値、最大値、及び最小値を求め、これを表 5 に示す。また、各手法を用いた場合のヒストグラムをそれぞれ図 6、及び図 7 に示す。横軸はコンセプト間の類似度であり、ピン幅は 0.1、縦軸はそのピン幅に対応する類似度を持つコンセプトの組の数の割合である。

これらの結果から、新手法によって似たコンセプトが形成されにくくなっていることを確認できた。コンセプト間の類似度の平均値、最大値、及び最小値はいずれも、旧手法よりも新手法の方が小さいため、新手法の方が似たコンセプトが形成されにくくなっていることがわかる。また、図 6 をみると以前の方法では、類似度が 0.7 以上のコンセプトの組が全体の 30% 程度であるのに対して、図 7 をみると新手法では、そのような類似したコンセプトの組は少なくなっていることがわかる。

6. 推薦実験

新手法によって得られたコンセプトを人の好みの表現に適用した場合の改良と有用性を調べるために、推薦実験を行った。旧手法と新手法のそれぞれで得られたコンセプトを用い、次の 6.1 節に示す二種類の推薦方法による、計四種類の推薦結果を被験者に評価してもらった。

6.1 推薦方法

本実験では、ユーザの好みに合う本として、次に示す二種類の本を推薦する。

- (A) ユーザが関心を持つコンセプトに関連強い本
- (B) ユーザの好みと似た特徴を持つ本

そのために、まず、「本の、各コンセプトとの関連強さ」及び「ユーザの、各コンセプトへの関心強さ」を求める。本とコンセプトとの関連度 $rel(b, co)$ は、その本 b に付けら

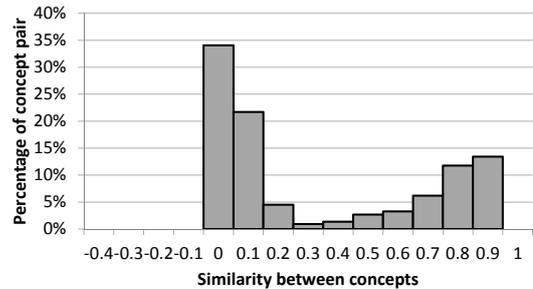


図 6 旧手法に関するヒストグラム

Fig. 6 A histogram about the previous method

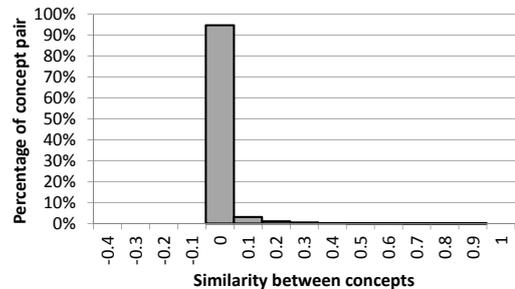


図 7 新手法に関するヒストグラム

Fig. 7 A histogram about the new method.

れている各タグ t_i とコンセプト co との間の関連度の平均値とする。また、ユーザのコンセプトへの関心度 $int(u, co)$ は、そのユーザ u が好きな本 b_i とコンセプト co との間の関連度の総和とする。

$$rel(b, co) = \frac{\sum_i rel(t_i, co)}{\text{(a size of tags attached to } b)}$$

$$int(u, co) = \sum_i rel(b_i, co)$$

そして、(A) のような本を推薦するために、各本 b の「ユーザ u が関心のあるコンセプトと b との間の関連強さ」 $pointA(b)$ を以下のように求め、この値が高い本を推薦する。また、(B) のような本を推薦するために、各本 b の「 b の特徴とユーザ u の好みとの間の類似度」 $pointB(b)$ を、コサイン類似度によって求め、この値が高い本を推薦する。

$$pointA(b) = \sum_i \{int(u, co_i) \times rel(b, co_i)\}$$

$$pointB(b) = \frac{\sum_i \{int(u, co_i) \times rel(b, co_i)\}}{\sqrt{\sum_i int(u, co_i)^2} \sqrt{\sum_i rel(b, co_i)^2}}$$

6.2 実験の内容と評価方法

あらかじめ、3 節で集めたデータ上の各本について、その本の特徴（各コンセプトとの関連度）を、旧手法で得られたコンセプトを用いた場合と、新手法で得られたコンセプトを用いた場合とで、それぞれ求めておく。次に、被験者には各自「好きな本」を列挙してもらい、それら本の特徴（各コンセプトとの関連度）から各被験者の好み（各コ

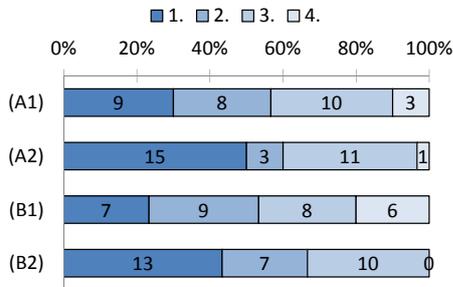


図 8 推薦実験の評価結果

Fig. 8 Results of the recommendation experiment

ンセプトへの関心度)を求める。被験者の「好きな本」が集めたデータ上にない場合は、その本に付けられているタグの情報をブックログから新たに得ることによって、その本の特徴を求める。求めた各被験者の好みに基づいて、6.1節の推薦方法を用い、以下四種類の推薦結果を各被験者に提示する。

(A1) *pointA* が高い上位 5 冊の本

(旧手法で得られたコンセプトを使用)

(A2) *pointA* が高い上位 5 冊の本

(新手法で得られたコンセプトを使用)

(B1) *pointB* が高い上位 5 冊の本

(旧手法で得られたコンセプトを使用)

(B2) *pointB* が高い上位 5 冊の本

(新手法で得られたコンセプトを使用)

そして被験者には、推薦された本それぞれに対する興味関心の有無を、以下の四段階で評価してもらう。

1. とても興味関心を持てる
2. 少し興味関心を持てる
3. あまり興味関心を持ってない
4. 全く興味関心を持ってない

6.3 実験結果

この推薦実験を被験者 6 名に対して実施した。(A1)~(B2) それぞれの推薦結果に対して、得られた 30 件 (6 名 × 5 冊分) の評価結果を図 8 に示す。

この推薦実験の結果から、新手法の改良と有用性が確認できた。推薦された本に対して興味関心があると評価した回答数は、(A1) が 17 件 (56.7%)、(A2) が 18 件 (60.0%)、(B1) が 16 件 (53.3%)、(B2) が 20 件 (66.7%) だった。旧手法で得られたコンセプトを用いた場合よりも、新手法で得られたコンセプトを用いた場合の精度が高かった ((A1) よりも (A2)、(B1) よりも (B2) の精度の方が高い)。また、これらの精度は他の研究と同等の精度である (例えば [7])。このため、人の好みの表現への、導出されたコンセプトの適用は有用であるといえる。

7. 結言

本研究では、人の好みを適切に表現するために、ユーザ達がアイテムに対して抱く印象を推測したものである「コンセプト」を用いることに着目した。以前の研究で提案した手法で実際に得られたコンセプトの妥当性を調べ、新たなコンセプトの形成方法を提案した。以前の手法からの改良のために、タグ間の類似度を再定義し、新たなタグのクラスタリング手法を提案し、タグとコンセプトとの間の関連度を拡張した。比較実験の結果、新たなコンセプトの形成方法の改良と有効性を確認した。また、推薦実験の結果から、新たな方法で得られるコンセプトを、人の好みの表現に適用することの改良と有用性を確認した。

参考文献

- [1] 田上道士, 山場久昭, 高塚佳代子, 岡崎直宣, 富田重幸: Serendipity を考慮した推薦システムの Folksonomy による実現とその評価, 宮崎大学工学部紀要 第 42 号, pp.313-318(2013)
- [2] Yamaba Hisaaki, Tanoue Michihito, Takatsuka Kayoko, Okazaki Naonobu, Tomita Shigeyuki: On a serendipity-oriented recommender system based on folksonomy, Artificial Life and Robotics, 18:89-94(2013)
- [3] 神鷹 敏弘: 推薦システムのアルゴリズム (2), 人工知能学会誌, vol.23, no.1, pp.89-103(2008)
- [4] J. Ben Schafer, Dan Frankowski, Jon Herlocker, Shilad Sen: Collaborative Filtering Recommender System, The Adaptive Web, LNCS4321, pp.291-324(2007)
- [5] Michael J. Pazzani, Daniel Billsus: Content-Based Recommendation Systems, The Adaptive Web, LNCS4321, pp.325-541(2007)
- [6] Philip L. Cahn: A non-invasive learning approach to building web user profiles, KDD.99 Workshop on Web Usage Analysis and User Profiling(1999)
- [7] 丹羽智志, 土肥拓生, 本位田真一: Folksonomy マイニングに基づく Web ページ推薦システム, 情報処理学会論文誌, Vol.47, No.5, pp1382-1392(2006)