

アイテムの主題分析に基づく協調フィルタリングシステムの推薦予測方式

宇田 隆 幸^{†1} 藤井 敦^{†2} 石川 徹 也^{†2}

我々は、協調フィルタリングに基づく情報推薦システムにおいて、推薦可能ユーザやコンテンツに対する網羅性を向上させる目的で擬似投票方式を研究してきた。擬似投票は、ユーザの評価したアイテムと内容的に類似するアイテムに対して、同じ評価値を付与する。しかし、アイテムに複数の主題が含まれる場合、ユーザの評価値がどの主題に向けられたものが不明である。この結果、「複数の主題を含むアイテムの場合、擬似投票の正確性は低下する」ことが予測される。そこで我々は、アイテムの主題（カテゴリ）が単数であるか複数であるかの判別を行い、主題数が単数である場合は従来と同様の擬似投票を行い、主題が複数である場合は擬似投票を抑制する方式を提案する。評価実験の結果、擬似投票における正確性、すなわちユーザの評価値とシステム評価値の類似性が向上した。

Recommendation prediction method in the collaborative filtering system based on the subject analysis of items

TAKAYUKI UDA,^{†1} ATSUSHI FUJII^{†2} and TETSUYA ISHIKAWA^{†2}

For recommender system based on the collaborative filtering, we have evaluated the Pseudo-Voting Method in order to improve the coverage for possible recommending users and contents. For the item which is contently similar to the item which the user evaluated, the Pseudo-Voting Method gives equal evaluated value. But, when multiple subjects are included for the item, it is uncertain the evaluated value of the user to which subject. As this result, "in case of the item including multiple subjects, the accuracy is low for the Pseudo-Voting" is estimated. Then, we distinguish "single or multiple the subject(category) which is included for the item". If a single subject of the item, we execute the traditional Pseudo-Voting. If a multiple subjects of the item, we propose the Suppressive Pseudo-Voting. The result of the evaluation experiment, the accuracy of Pseudo-Voting Method, namely the similarity of user's evaluated value and Pseudo-Voting value was improved.

1. はじめに

1.1 情報推薦システムの問題点と本研究の目的

Web マイニングの応用技術として、情報推薦システムが研究されてきた^{1), 2), 3), 4), 5)}。しかし、研究において評価実験が行われているシステムや、Amazon.com⁶⁾、⁷⁾などに代表される電子商取引サイトが中心であり、ユーザ向けの情報を対象にしているシステムは未だない。また、WWW は世界中の情報を対象にするが、現実社会におけるユーザの行動範囲に的を絞った情報、すなわち地域情報を対象としたシステムは、著者らが知る限りない。

先行研究では、情報推薦システムのアルゴリズムの一つである協調フィルタリングアルゴリズムが中心に扱われてきた（表 1、表 2 参照）。協調フィルタリングの問題点として、アイテムおよびユーザ数が大規模になるほど、評価値疎ら問題（Sparsity problem）や再生起問題（Recurring startup problem）等が指摘されている。この 2 点の問題を解消する方式として、擬似投票方式が提案された¹⁰⁾。擬似投票方式は、仮説「あるアイテムを好むユーザは、文書的に類似する別のアイテムも同じように好む」を前提に成立する。仮説の妥当性も評価実験によって証明されており¹⁰⁾、2 点の問題もある程度解消されている。

擬似投票方式で利用されるアルゴリズムは、アイテムに含まれる語の出現頻度を利用したアイテム間の類似度計算を行うことで類似するアイテムを探しだし、そのアイテムに対して、ユーザが評価した投票値と同じ値をシステムが擬似的に投票する。そして、ユーザによる評価値と擬似投票値を参考にして、他のユーザ

†1 筑波大学・図書館情報メディア研究科・博士後期課程
Graduate School of Library, Information and Media Studies,
University of Tsukuba
(株式会社ネオジェイエスケー NEO JSK Corporation)
†2 筑波大学・図書館情報メディア研究科
Graduate School of Library, Information and Media Studies,
University of Tsukuba

に対して推薦アイテムを提示する。

協調フィルタリングに基づくシステムは、アイテムの内容解析をせず、他のユーザからの評価（アイテム-ユーザ評価マトリクス）を用いて推薦アイテムを決定する。しかし、ユーザの評価は、実際にはアイテム全体に対する評価ではなく、アイテムの一部分（真に必要な主題）に対する評価であるはずである。すなわち、複数の主題（カテゴリ）を含むアイテムに対するユーザの評価は、特定の主題に対する評価であるはずである。そこで、擬似投票方式において、アイテムの推薦スコア予測計算の正確性を確保するために、アイテム内の主題判別が必要になる。

1.2 研究の範囲

上述の背景から、アイテムの推薦スコア予測計算における正確性を高める目的で、本研究では「Web ページによる実際の地域情報アイテム」を対象として、ユーザによって評価されたアイテムが複数の主題を含むかどうかを解析し、アイテムが複数の主題を含む場合と否とに判別し、推薦スコアの予測計算に利用する方式を研究する。この判別結果を利用して、複数の主題を含むアイテムについては、ユーザ間の嗜好の類似度計算における相関度の重みを減少させたり、推薦スコアを減少させたりすることで、推薦スコア予測計算における正確性を高める。

実験は、アイテム間の類似度計算において好成绩を記録した Dice 係数 (Dice formulation) と、アイテムの一部が重複する場合に高い類似性を示す部分一致公式 (Overlap formulation) との差分を用いて、主題が単数であるか複数であるかを判別するための閾値を求める。そして、この判別結果を利用するため、推薦スコア予測計算式における重みの増減方法について最適値を探す。

以下 2 章で関連研究を整理し、3 章で協調フィルタリング方式の問題点を示し、4 章で本研究システムについて示し、5 章で評価実験結果を示し、6 章で結論を示す。

2. 関連研究

情報推薦システムは、推薦情報を予測計算するときに着目するアイテム（情報）により分類すると表 1 のように整理できる。この中で「狭義の情報推薦方式」は、ユーザの行動履歴を参考にして推薦情報を生成するので、システム稼働のために、手動で行う準備作業が不用である点、ユーザ数やアイテム数の増加に追従できる点、意外性のあるアイテムを推薦できる点、推薦アイテムの予測計算を自動で行える点、

表 1 情報推薦システムの分類

Table 1 Classification of existing recommender systems.

	方式	着目する情報
(1)	チェックボックス方式	ユーザの登録した嗜好情報
(2)	ルールベース方式	サイト運営者の登録した推薦ルール
(3)	狭義の情報推薦方式 ^{1), 4), 5), 8), 9), 10), 11)}	ユーザの行動履歴

表 2 狭義の情報推薦方式の分類

Table 2 Classification of recommender system methods.

	方式
(1)	内容ベースフィルタ方式 ^{1), 8)}
(2)	協調フィルタリング方式
	相関係数法 ^{5), 10), 11)} クラスタモデル ⁹⁾
(3)	ハイブリッドフィルタ方式 ^{1), 4)}

ユーザはクリックを除くキー入力が不用である点において、他の方式よりも優位である¹²⁾。

「狭義の情報推薦方式」における推薦エンジンでは、フィルタリングアルゴリズムが用いられている。推薦情報を自動生成するとき、着目する情報や推薦スコア計算方式に着目すると、「狭義の情報推薦方式」は、表 2 のように整理できる。この中で「相関係数法を用いた協調フィルタ方式」は、システム側の事由で予測精度を低下させない点、意外性のあるアイテムを推薦できる点、ユーザの嗜好特性や情報特性に影響されない点において、他の方式より優位である⁶⁾。そのため、本研究では、相関係数法を用いた協調フィルタリング方式（以下、協調フィルタリング）をベースにしている。

3. 協調フィルタリング方式の問題点

この節では、本研究で提案するシステムを構築する上で基礎となる情報推薦システムと協調フィルタリングアルゴリズムについて述べる。そして、協調フィルタリングの推薦スコア予測計算方式における問題点を解消させるための研究課題について述べる。

3.1 協調フィルタリングアルゴリズム

ユーザ u_j が評価投票したアイテムの集合を $I_{u_j} \in I$ とする。ユーザ u_j が評価投票したアイテム k に対する評価投票値を $r_{j,k}$ とする。ユーザに対して推薦を行う場合、協調フィルタリングを行うとは、ユーザ $\{u_1, u_2, \dots, u_m\}$ のアイテム $\{i_1, i_2, \dots, i_n\}$ に対する評価 $r_{j,k}$ を参考にして、ユーザ u_j が未評価のアイテム、すなわ

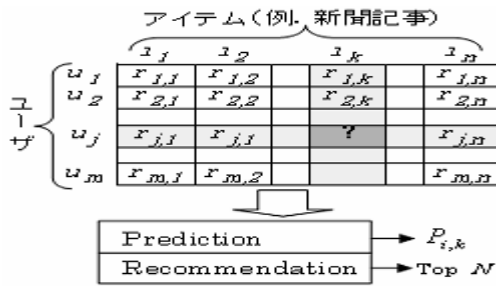


図 1 協調フィルタリングの概念図
Fig. 1 Overview of collaboration filtering.

ちアイテム $i_j \notin I_{u_j}$ に対する評価値 $r_{j,k}$ を予測 ($P_{j,k}$) することである。ただし、ユーザ u_j に対して推薦されるアイテムは予測値の中で嗜好度の高い上位 N アイテムである。代表的な $P_{i,k}$ の算出法は、式 1 である。

$$P_{i,k} = \bar{r}_i + \frac{\sum_m \text{sim}(u_i, u_m) \cdot (r_{m,k} - \bar{r}_m)}{\sum_m \text{sim}(u_i, u_m)} \quad (1)$$

ここで、 $P_{i,k}$ はユーザ u_i に対するアイテム k への嗜好度合いの予測値、 \bar{r}_i はユーザ u_i が行った全ての評価投票の平均投票値、 m はアイテム k へ評価投票を行ったユーザ集合、 $\text{sim}(u_i, u_m)$ はユーザ u_i とユーザ u_m の嗜好度合いの類似度 (式 2)、 $r_{m,k}$ はユーザ u_m がアイテム k へ行った評価投票値、 \bar{r}_m はユーザ u_m が行った全ての評価投票の平均投票値とする。

ユーザ間の類似度 $\text{sim}(u_i, u_m)$ は式 2 を用いる。

$$\text{sim}(u_i, u_m) = \frac{\sum_j (r_{i,j} - \bar{r}_i) \cdot (r_{m,j} - \bar{r}_m)}{\sqrt{\sum_j (r_{i,j} - \bar{r}_i)^2 \cdot (r_{m,j} - \bar{r}_m)^2}} \quad (2)$$

ここで、 j はユーザ u_i とユーザ u_m の双方が評価投票を行っているアイテム集合、 $r_{i,j}$ はユーザ u_i がアイテム j へ行った評価投票値、 $r_{m,j}$ はユーザ u_m がアイテム j へ行った評価投票値である。

3.2 擬似投票

擬似投票は、「あるアイテムを嗜好するユーザは、内容的に類似する他のアイテムも同じように嗜好する」という仮説に基づいており、過去の評価実験によって仮説の妥当性を証明している^{10), 11)}。擬似投票アルゴリズムを以下に示す。

```

擬似投票(UserId, ItemId_U, RealVoteVal) {
  For each(システムに蓄積されている全アイテム: ItemId_S) {
    If (類似判定(ItemId_U, ItemId_S, 類似閾値) == 類似
      && ユーザが過去に評価していない(ItemId_S) == YES) {
      擬似投票値記録(ItemId_S, RealVoteVal);
    }
  }
}

```

擬似投票処理を実施すると、図 2 に示すように「アイテム - ユーザ」評価マトリクスの未評価セルに評価値を設定することができる。未評価セルが減少するこ

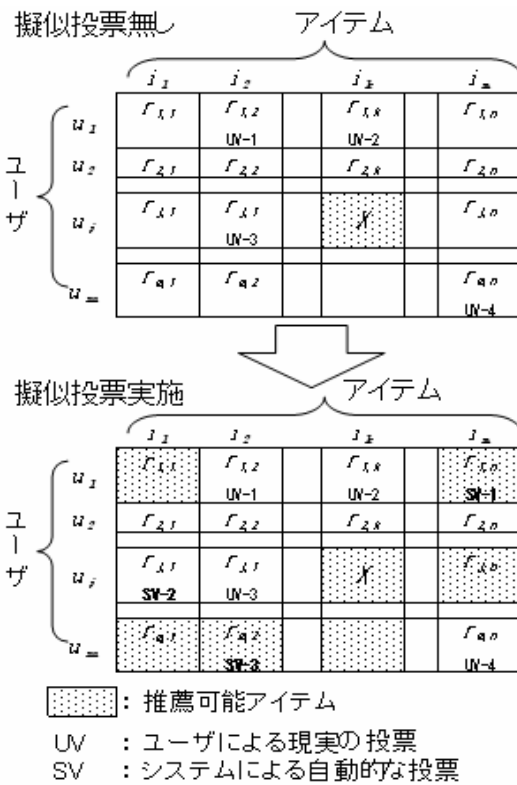


図 2 擬似投票の方式
Fig 2 Algorithm of pseudo voting method.

とで、協調フィルタリングに基づく推薦予測計算において、推薦可能ユーザや推薦可能アイテムに対する網羅性が高まる。

3.3 問題点

擬似投票は前述の通り、ユーザが評価したアイテムと内容的に類似する未評価アイテムを探し、そのアイテムの評価値格納用のセルに対し、ユーザが評価した値と同値を設定する。

しかし、図 3(a), (b) に示すように、単一の主題から構成されるアイテムと、複数の主題から構成されるアイテムとが存在する。図 3(a)の場合、ユーザの評価値はページ全体に対する評価である。しかし、図 3(b)の場合、ユーザの評価値は、ページ全体、主題 1 あるいは主題 2 に対する評価と考えられる。このような理由から、複数の主題を含むアイテムにユーザが投票した値を利用すると、擬似投票の正確性が低下する。その結果、推薦予測計算の正確性も低下する。

サンケイリビング新聞社が提供している地域情報サイト「えるこみ」。同サイトでは、女性のために生活情報が提供されている。評価実験で使用したアイテムの種類は、イベント・教養&カルチャー・生涯学習・グルメなどの地域情報である。
<http://www.lcomi.ne.jp/>

図 3(a) . アイテムの例 (単一主題)
Fig 3(a) Example of item for single subject.

図 3(b) . アイテムの例 (複数主題)
Fig 3(a) Example of item for multiple subject.

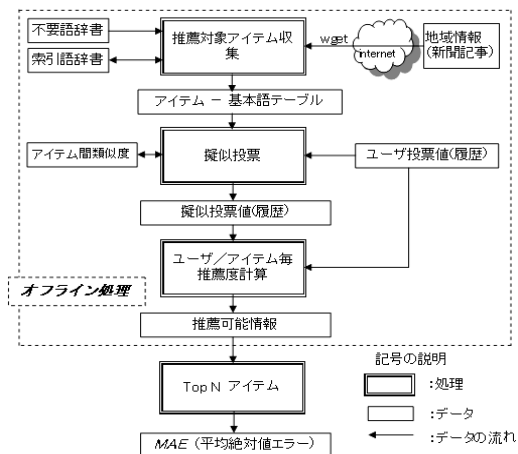


図 4 . システムの全体図
Fig 4 System overview.

4. 研究システム

図 4 は、実験を行う目的で構築したシステムの全体図である。擬似投票処理は、ユーザ評価履歴を入力として、アイテムペアの類似度テーブルを参照し、類似度の高いアイテムに対して擬似投票を行い、結果をアイテム - ユーザ評価マトリクスに記録する。

本研究では、仮説「複数の主題を含むアイテムの場合、擬似投票は正確性が低くなる」の妥当性を検証する。

4.1 アイテムの主題分析方法

上述の問題点の解消を目的に、アイテムの主題の数に着目した推薦スコア予測計算方式を提案する。ここでは、アイテムの主題数が複数であるか単数であるかを判別する方式を示す。

擬似投票は、ユーザの評価したアイテムと内容的に類似するアイテムに対して、同じ評価値を付与する。そのため、擬似投票を行うためには、2つのアイテムに対する内容的な類似性を計算する。

本研究では、索引語の重み付けは IDF を用いないで TF のみで求める。アイテム間の類似度計算方式は、式(3)と式(4)を用いる。

Dice 係数方式(Dice formulation)

$$S_{u,s} = \frac{2 \sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}{W_u^2 + W_s^2} \quad (3)$$

部分一致公式(Overlap formulation)

$$S_{u,s} = \frac{\sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}{\min(W_u^2, W_s^2)} \quad (4)$$

ここで、 $S_{u,s}$ はアイテム u とアイテム s の類似度、 T

は異なり語の集合, $t \in T_{u,t}$ における t はアイテム u とアイテム s の両方に含まれる語である. $w_{u,t}$ と $w_{s,t}$ はアイテム u またはアイテム s に含まれる語 t の重みとする. ここで, $w_{u,t} = 1 + \ln(f_{u,t})$, $w_{s,t} = 1 + \ln(f_{s,t})$ とし, $f_{u,t}$ と $f_{s,t}$ は, アイテム u またはアイテム s における語 t の出現頻度である. アイテムの長さは, $W_u = \sqrt{\sum_{t \in T_u} w_{u,t}^2}$, $W_s = \sqrt{\sum_{t \in T_s} w_{s,t}^2}$ とする.

主題分析は, 式(3)と式(4)の計算結果の差分に基づいて, 一方のアイテムに複数の主題が含まれているか否かを判別する.

4.2 主題分析結果による擬似投票

擬似投票は, 3.2 節で記述した方法と同じである. ただし, 擬似投票の対象は, 内容的に類似するアイテムが対象という条件に加え, 主題分析で複数の主題が含まれていないアイテムペアとする.

4.3 主題分析結果による協調フィルタリング

主題分析結果は, 擬似投票段階で反映されている. 従って, 協調フィルタリングは, 式(1)と式(2)を用いて行う.

5. 評価実験

5.1 評価実験の方針

(1) 主題分析の方針

擬似投票の正確性を高めるために行う主題分析は, 式(3)と式(4)の計算結果に基づいて, アイテムの主題数が複数であるか単数であるかを判別する. 以下, 式(3)と式(4)の計算結果に着目した主題分析方法の方針を示す.

(A) 式(3)・式(4)とも類似性が低い

2 つのアイテム間の類似性は低いので, 擬似投票の対象外となる. この場合, アイテムの主題数を分析する必要性がない.

(B) 式(3)の類似度が高く, 式(4)の類似度が低い

Dice 係数方式の類似性判定結果は高い. しかし, 部分一致公式を用いた類似性判定結果は低い. 式(3)と式(4)の違いは分母にある. 式(3)と式(4)の分母を比較すると, 式(4)の分母は, 式(3)の分母より小さな値となる. また, W_u と W_s はいずれも 1 以上であるから, 式(3)の分子に係数 2 が乗算されていても, 式(3)の計算結果が式(4)の計算結果より大きく (類似度が高く) なることはない.

(C) 式(3)・式(4)とも類似性が高い

2 つのアイテム間の類似性は高く, 擬似投票の対象である. 両式の計算結果に違いが認められない場合とは, 両アイテムの長さ, すなわち W_u と W_s が近似している. 両アイテムの類似性が高い場

合, 両アイテムは同じ主題に関する内容であり, 類似性が低い場合は, 別々の主題に関する内容である. この場合, 主題分析の必要はなく, 類似性のみを検討すればよい.

(D) 式(3)の類似度が低く, 式(4)の類似度が高い

Dice 係数方式の類似性判定結果は低い. しかし, Overlap Formulation (以下, 部分一致公式) の類似性判定結果は高い. この場合, アイテム全体の類似性は低いものの, アイテムの一部 (語数の少ない方のアイテムが基準) が一致していることを示している. このことから, 語数を多く含むアイテムには, 主題が複数含まれている可能性が高い.

アイテムの主題数が単数であると判別する条件は, 「式(3)と式(4)の差分が小さいこと」である. 本研究では, 30 アイテムペアを人手で観察して, 好成績を残した閾値を決定する.

(2) 擬似投票の方針

擬似投票を実施するアイテムペアは,

条件 1: アイテムが類似していること

条件 2: 複数の主題が含まれていないこと

を条件とする. 条件 1 は, 我々が過去に行った実験で求めた最適値を用いる¹¹⁾ 条件 2 は, 値を変化させて, 擬似投票の正確性を高める条件を観察し, 閾値を決定する.

5.2 評価実験データ

研究用データは, 「えるこみ」に蓄積されているアイテムのうち, 2002 年 10 月 27 日 ~ 2003 年 2 月 15 日 (12 週間) の間に掲載されたアイテムを対象として, その内 6% を標本化した. アイテムの標本数は, 2,424 アイテムである.

アイテムページの例は, 図 3 を参照. ユーザの評価値は, 「とても」(3 点), 「まあまあ」(2 点), 「いまいち」(1 点) とした. 研究用ユーザ評価値投票履歴データは, 上記 12 週間においてユーザが直接評価投票した履歴を標本化した. 標本数は 1,275 投票であり, 異なりユーザ数は 756 人である. なお, 「同一ユーザによる同一アイテムへの評価投票」は, 第一印象を重視して, 初回の評価投票値のみを利用している.

なお, 実験データを観察し, 主題数分析を行ったところ, 下記 4 件のフレーズは多くの新聞記事に同じように出現する. そして, Dice 計数による類似度は低く, 部分一致公式の類似度が高いという結果が引き起こされる可能性が高いため, 主題分析 (閾値) の決定処理では取り除いている.

分析対象外のフレーズ: 「日付情報, 地域を特定

する情報、電話やその他のアクセス情報」。

5.3 主題分析結果

類似するユーザの評価したアイテムと類似するアイテムに対して、システムは擬似投票を実施する。従って、擬似投票に先立ち、システムは2つのアイテム間の類似度を計算する。この時点で主題分析結果を、式(3)と式(4)の差分として求める。

アイテムに複数の主題が存在すると判別するための閾値は、式(3)と式(4)の差分の絶対値が0.8を超える場合とした。この閾値を満足するアイテムペア30個を手作業で分類したところ、28ペア(93.4%)が正解で、2ペア(6.6%)が不正解であった。

5.4 主題分析結果の効果

表3は、投票期間別の擬似投票の結果を示している。表3における、期間IDは投票期間を示しておりID=1は実験開始後～1週間、ID=12は実験開始後～12週間のユーザ投票を使用した。擬似投票数は、当該期間における擬似投票の件数を示している。MAE(mean absolute error)は、先行研究においてシステム評価方式として提案されている「ユーザによる評価投票値と予測値間の平均絶対値エラー」であり、式(5)で計算した。MAEがゼロに近づくほど、ユーザによるアイテムへの評価値とシステムによる擬似投票値の嗜好パターンが類似していることを意味する。

$$MAE = \frac{\sum_{j \in U} \sum_{k \in T_{u_j, s_j}} \frac{|r_{j,k} - sv_{j,k}|}{N_k}}{N_j} \quad (5)$$

式(5)において、 j はアイテムへの評価投票を行ったユーザの集合で、 N_j は集合 j の要素数(異なりユーザ数)、 k はユーザ j が評価したアイテムとシステムがユーザ j のために擬似投票したアイテムの積集合、 $r_{j,k}$ はユーザ j がアイテム k に与えた評価値で、 $sv_{j,k}$ はシステムがユーザ j のアイテム k に対して行った擬似投票値の平均値、 N_k は集合 k の要素数(アイテム数)である。なお、計算に先立ち、 $r_{j,k}$ と $sv_{j,k}$ は、-1.0～+1.0の値域に基準化した。

ここでは、擬似投票における主題分析の効果を測定した。主題分析を行わない場合の擬似投票は、式(3)の計算結果が0.9以上を類似するアイテムペアとした。また、主題分析を擬似投票に反映する条件は、式(3)の計算結果が0.9以上で、かつ、式(3)と式(4)の差分が0.1未満を対象とした。

この結果、主題分析結果の効果は、表3に示した通り、MAEが改善している。

表3. 擬似投票の結果

Table 3 The Pseudo-Voting result.

期間 ID	ユーザ投票数	主題分析なし		主題分析実施	
		擬似投票数	MAE	擬似投票数	MAE
1	91	44,446	0.111	24,765	0.000
2	251	121,912	0.094	68,622	0.076
4	602	277,517	0.106	161,020	0.065
8	1,256	556,647	0.169	323,529	0.108
12	2,197	947,486	0.183	550,914	0.125

表4. 推薦予測結果

Table 4 The recommendation prediction result.

期間 ID	主題分析なしの MAE	主題分析実施の MAE
1	0.760	0.752
2	0.650	0.647
4	0.526	0.531
8	0.494	0.484
12	0.468	0.449

表5. 表4における誤差計算方法

Table 5 Error calculation rule in Tble 4.

No	ユーザ評価	システム評価	妥当性	計算対象
1	1:とても好き	1:上位100件	正解	対象
2	1:とても好き	上位100外	不正解	対象外
3	0:まあまあ	1:上位100件	-	対象外
4	0:まあまあ	上位100外	-	対象外
5	-1:いまいち	1:上位100件	不正解	対象
6	-1:いまいち	上位100外	正解	対象外

5.5 協調フィルタリングの推薦予測結果

ユーザ投票値と擬似投票値の両方に対して式(1)にて協調フィルタに基づく推薦スコアの予測計算を行い、ユーザ評価値システム評価値間の誤差を表4に示した。

表4のMAEは、式(6)で計算した。計算対象と評価に対する変数値を、表5に示した。

$$MAE = \frac{\sum_{i=1}^N | \text{ユーザ評価}_i - \text{システム評価}_i |}{N} \quad (6)$$

ここで N は計算対象の組み合わせ数である。

この結果、主題分析を行うことで、表4に示した通り、MAEがわずかながら改善した。

5.6 考察

(1) 主題分析

主題分析の正解率は93.4%であり、本方式は高い精度の方式であるといえる。不正解となったアイテムは、

アイテムの主題に関する内容記述が長文であるか短文であるかの違いである。

(2) 主題分析結果に基づく擬似投票結果

擬似投票の結果は、表 3 の通り主題分析の有無にかかわらず同程度の正確性であった。これは、擬似投票を実施したアイテムペアの一方に、主題を複数含むアイテムの割合が 8%であったためである。このため、主題分析によって擬似投票の正確性は向上するものの、擬似投票全体に対する効果はわずかである。

(3) 協調フィルタリング予測の正確性

協調フィルタリング予測の正確性についての結果は、表 4 の通りである。主題分析結果を反映した擬似投票件数が少ないため、協調フィルタリングに基づくアイテムの推薦スコアに大きな変化がない。そのため、システムの推薦する上位 100 件のアイテムもほとんど同じアイテムが推薦された。この結果、推薦アイテムの正確性は、主題分析の実施有無にかかわらず、差が生じなかった。

なお、ユーザ毎アイテム毎の推薦スコア計算後の並べ替え処理量を減少させる目的で、式(1)の値が推薦スコア上位 100 件に含まれないことが明らかな場合は、そのアイテムに対する推薦は行わなかった。このため、MAEの計算は推薦スコアが低いアイテムについて計算を行えないため、他の先行研究³⁾と比べ低い値を示している。

6. 結論

主題分析において、一方のアイテムに複数の主題が含まれているか否かの判定は、Dice 係数の計算結果と部分一致公式の計算結果の差分に着目することで、判別することができた。

また、評価実験から、仮説「複数の主題を含むアイテムの場合、擬似投票は正確性が低い」の妥当性がわずかながら観測できた。このことから、当仮説の妥当性を示すことができた。

主題分析結果によって推薦予測精度の正確性が向上するためには、(1)主題分析の有用性が高く、かつ(2)その有用性を観測できるほどに高い割合で複数の主題が含まれるアイテムが必要である。本研究では、実験結果から、(1)の有用性わずかに示すことができた。しかし、実験データが地域新聞の記事であるため、(2)と背反するアイテムを取り扱ったため、(1)に対する大きな効果を観測できなかった。

今後は、主題分析とは別の方式で、情報推薦システムにおける推薦アイテムの正確性を向上させるための研究を行う。

謝辞 評価実験用データを貸与いただきましたサンケイリビング新聞社の皆様に感謝します。

参考文献

- 1) Balabanovic, M., Shoham, Y. : Fab: Content-Based, Collaborative Recommendation. In CACM. Vol.40, No.3 pp.66-72 (1997).
- 2) Hirooka, Y., Terano, T. and Otsuka, Y. : Extending Content-Based Recommendation by Order-Matching and Cross-Matching Methods: In Bauknecht, K. et al. eds.: Electronic Commerce and Web Technologies - 1st Int. Conf. EC-Web 2000, Springer Verlag, LNCS 1875, pp.177-190, (2000).
- 3) Polcicova, G., Navrat, P.: Semantic Similarity in Content-based Filtering : In Advances in Databases and Information Systems, 6th East European Conference, ADBIS 2002, Bratislava, Slovakia, September 8-11, 2002, Proceedings. Lecture Notes in Computer Science 2435 Springer 2002, ISBN 3-540-44138-7, pp.80-85(2002).
- 4) Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. : Combining Content-Based and Collaborative Filters in an Online Newspaper : in Proc. Recommender Systems Workshop at ACM SIGIR(1999).
- 5) Resnick, P., Iacovou, N., Suchak, M., Bergström, P. and Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews, Proc. of the CSCW 1994, New York, pp.175-186, ACM (1994).
- 6) Linden, G., Smith, B., and York, J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering: in Proc. of February 2003 issue of IEEE Internet Computing (2003).
<http://dsonline.computer.org/0301/d/w1lind.htm>
- 7) Schafer, J.B., Konstan, J.A. and Reidl, J. : E-Commerce Recommendation Applications: in Proc. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, pp.115-153(2001).
- 8) Basu C., Hirsh, H. and Cohen, W. : Cohen. Recommendation as Classification: Using Social and Content-based Information in Recommendation. Proc. Recommender System Workshop AAAI 1998, pp.11-15, 1998.
- 9) L. H. Ungar and D. P. Foster. Clustering Methods for Collaborative Filtering. Proc. Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence, pp.114-129, 1998.
- 10) 宇田隆幸, 藤井敦, 石川徹也.: ユーザ投票と情報アイテム間類似度を併用した情報推薦システム: 擬似投票方式の提案と評価: 情報処理学会研究報告, 2004-DBS-132, pp.105-112, 2004.
- 11) 宇田隆幸, 藤井敦, 石川徹也.: 協調フィルタリングに基づく情報推薦システムにおける擬似投票方式: 言語処理学会第 10 回年次大会発表論文集, pp.381-384,

2004.

- 12) 寺野隆雄, “情報推薦システム,” eビジネスの理論と応用, 北村泰彦, 山田誠二(編), 第3章, 東京電機大学出版局, 東京, (2003).