DBSJ Journal, Vol.12, No.1
June 2013

閲覧履歴グラフに基づく正則化リンク解析を用いたロバスト推薦

Robust Recommendations using Regularized Link Analysis of Browsing Behavior Graphs

内藤 慎也 ♥　江口 浩二 ♦

Shinya NAITO　Koji EGUCHI

近年 Web で提供されるデータの増加に伴い，情報推薦技術の高度化への要求が高まりつつある．とりわけ本論文では比較的長いテキストデータが対応づけられたアイテムを対象とした情報推薦に着目する．この目的のもと，Co-HITS アルゴリズムに基づき，ユーザとアイテムからなる 2 部グラフに対して相互強化によるリンク解析を行うことによって情報推薦を実現する．本手法は，グラフの構造とアイテムの内容情報を，正則化によって統合するものである．実験では，Web ニュースを対象アイテムとし，閲覧履歴に基づいて推薦された上位 N 件のアイテムリストを評価する．この実験により，情報推薦においてしばしば問題となるデータスパースネス問題が発生する状況において提案手法が複数のベースライン手法より有効に機能することを示す．

Recently, there has been a growing need for more sophisticated recommendation techniques with an increase in the amount of data available on the Web. In this study, we especially focus on recommending items associated with long text, and aim at achieving this using a method of link analysis of a user-item bipartite graph in a framework of mutual reinforcement based on Co-HITS algorithm. This method integrates the graph structure and the content of items by regularization. We demonstrate through experiments on top-N item recommendations of Web news that the proposed method outperformed several baseline methods in a situation where only a small amount of browsing behavior is observed.

♥ 非会員　神戸大学大学院システム情報学研究科
s-naito@cs25.scitec.kobe-u.ac.jp
♦ 正会員　神戸大学大学院システム情報学研究科
eguchi@port.cs.kobe-u.ac.jp

1. はじめに

近年 Web データの増加に伴い，情報推薦技術の高度化への要求が高まりつつある．典型的な情報推薦の問題は，ユーザに対してその好みに合致するアイテムを一定件数だけ提示する上位 N 推薦と，ユーザの特定アイテムに対する評価値を予測する評価値予測である．この情報推薦の手法としては様々なものが提案されているが，代表的な手法としてはコンテンツベースフィルタリングと協調フィルタリングの 2 つがある [1]．ところで，協調フィルタリングの手法を用いて情報推薦を行う際，ユーザやアイテムについての履歴情報が少ない場合に十分な推薦ができないというデータスパースネス問題が生じることがある．これは協調フィルタリングが，主として推薦対象のユーザがこれまでにどのような行動をとっているか，または推薦対象に類似したユーザがどのような行動をとっているかという情報を利用しているためである [2, 3]．また，協調フィルタリングで発生する問題としてコールドスタート問題 [4] がよく知られている．これは履歴情報がほとんどない新着ユーザ（コールドスタートユーザ）や新着アイテム（コールドスタートアイテム）に対して発生するものであり，データスパースネス問題の一種として捉えることができる．本論文では比較的長いテキストデータが対応づけられたアイテムを対象とし，閲覧履歴や購入履歴などのユーザ行動履歴に基づいた情報推薦を考える．このような目的のもと，我々はコールドスタート問題を含むデータスパースネス問題に注目し，これを解決するために Co-HITS アルゴリズム [5] を導入する．Co-HITS アルゴリズムは元々はクエリ推薦を目的とした開発されたものであり，本論文で目的とする情報推薦にそのまま適用するのは適切でないため，種々の観点から修正を行う．本論文では，上位 N 推薦に関する実験に基づいて，コンテンツベースフィルタリング及び HITS アルゴリズム [6] を用いた協調フィルタリングと比較し，提案手法によってデータスパースネス問題に対してよりロバストな推薦が行えることを示す．

2. Co-HITS アルゴリズム

本節では Co-HITS アルゴリズム [5] の概要について説明する．

まず，いくつかの定義について述べる．2 部グラフ $G = (U \cup V, E)$ を考えるとき，各ノードは互いに素な 2 つの集合 U，V に分割でき，辺は一方の集合のノードから他方の集合のノードを結び，同一集合内のノードを結ぶ辺は存在しない．ここで，U と V をそれぞれ $U = \{u_1, u_2, \ldots, u_m\}$，$V = \{v_1, v_2, \ldots, v_n\}$ と表現する．この 2 部グラフ上のランダムウォークを考える．U に含まれるノード u_i と V に含まれるノード v_j を結ぶ辺が存在するとき，u_i と v_j の間の遷移確率を w_{ij}^{uv} 及び w_{ji}^{vu} で示すことができる．2 ノード間に辺が存在しないときは遷移確率が 0 となる．また，あるノードに着目したとき，他ノードへの遷移確率の和と他ノードからの遷移確率の和はそれぞれ 1 となる．さらに，実際には辺が存在しない同一集合内のノード間について，隠れ遷移確率 w_{ij}^{uu} 及び w_{ij}^{vv} を考えることができる．例えば，u_i から u_j への隠れ遷移確率 w_{ij}^{uu} は，$w_{ij}^{uu} = \sum_{k \in V} w_{ik}^{uv} w_{kj}^{vu}$ によ

り得られる．これ以降，U から V への遷移については遷移行列 $W^{uv} \in \mathbb{R}^{m \times n}$ で表す．ここで，W^{uv} は (i,j) 成分が u_i から v_j への遷移確率 w_{ij}^{uv} を示す．同様に，$W^{vu} \in \mathbb{R}^{n \times m}$ によって V から U への遷移を表す．また，$W^{uu} \in \mathbb{R}^{m \times m}$ と $W^{vv} \in \mathbb{R}^{n \times n}$ によって U と V それぞれの内部における隠れ遷移行列を表す．

Deng らの Co-HITS アルゴリズム [5] は，まず 2 部グラフの各ノードに対応づけられたテキストデータに基づいてクエリに対する各ノードの適合度の初期値（以下，初期適合度）を与え，次に各ノードから他ノードへの遷移確率に基づいて適合度の更新処理を繰り返し行う．それによって，各ノードに対応づけられたテキストデータの内容情報及びグラフの構造の両方を考慮した適合度に更新することが可能となる．Co-HITS アルゴリズムによる u_i の適合度 x_i と v_k の適合度 y_k は，それぞれ式 (1)，(2) のように定義できる．

$$x_i = (1-\lambda_u)x_i^0 + \lambda_u \sum_{k \in V} w_{ki}^{vu} y_k, \quad (1)$$

$$y_k = (1-\lambda_v)y_k^0 + \lambda_v \sum_{j \in U} w_{jk}^{uv} x_j \quad (2)$$

ここで，$\lambda_u \in [0,1]$ と $\lambda_v \in [0,1]$ はいずれも各ノードの内容情報とグラフの構造の重みを決定するパラメータで，x_i^0 と y_k^0 は u_i と v_k それぞれの初期適合度である．前述の反復操作が繰り返されると，適合度は各ノードの初期適合度及びその近傍の状況に基づいて，ある値に収束する．この反復過程において，Co-HITS アルゴリズムでは，初期適合度やグラフ構造に基づく正則化項を導入することにより，種々の制約を考慮することができる．以下にその詳細について説明する．

2 部グラフのノード集合 U について，コスト関数 R_1 を次のように定義できる．

$$R_1 = \frac{1}{2} \sum_{i,j \in U} w_{ij}^{uu} \left\| \frac{x_i}{\sqrt{d_{ii}}} - \frac{x_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i \in U} \| x_i - x_i^0 \|^2 \quad (3)$$

ここで $\mu > 0$ は正則化パラメータであり，d_{ii} は対角行列 D の対角要素 $d_{ii} = \sum_j w_{ij}$ を示し，正規化のために用いる．コスト関数の第 1 項はグラフにおける適合度の大域的一貫性を与え，第 2 項は初期適合度による制約を与える．それらの重みはパラメータ μ で調整することができる．同様に，V に関連したコスト関数 R_2 は次のように表される．

$$R_2 = \frac{1}{2} \sum_{i,j \in V} w_{ij}^{vv} \left\| \frac{y_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i \in V} \| y_i - y_i^0 \|^2 \quad (4)$$

R_1 と R_2 では U と V それぞれのグループ内における隠れリンクに基づいた一貫性が定められたが，U と V の間の直接的なリンクを考慮するコスト関数 R_3 が次のように定義できる．

$$R_3 = \frac{1}{2} \sum_{i \in U, j \in V} w_{ij}^{uv} \left\| \frac{x_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right\|^2 + \frac{1}{2} \sum_{j \in V, i \in U} w_{ji}^{vu} \left\| \frac{y_j}{\sqrt{d_{jj}}} - \frac{x_i}{\sqrt{d_{ii}}} \right\|^2 \quad (5)$$

R_3 は，U と V にまたがる強連結なノード対において適合度に大きな差があった場合にペナルティーを課すものである．

最終的に，U と V の双方に対応づけられたコスト関数 R は次のように定義される．

$$R = \lambda_\gamma (R_1 + \alpha R_2) + (1-\lambda_\gamma) R_3 \quad (6)$$

ここで，$\alpha > 0$，$\lambda_\gamma \in [0,1]$ である．コスト関数 R を最小化することで，ノードの内容情報とグラフの構造を考慮した適合度が得られる．各要素が対応するノードの適合度を表す適合度ベクトルを $F = (f_1, \ldots, f_{m+n})^T = (x_1, \ldots, x_m, y_1, \ldots, y_n)^T$ とするとき，最小化問題 $\min_F R$ は次式に書き換えられる．

$$\min_F \ \frac{1}{2} \sum_{i,j=1}^{m+n} w_{ij} \left\| \frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i=1}^{m+n} \| f_i - f_i^0 \|^2$$

$$s.t. \quad W = \begin{bmatrix} W^{uu} & \beta \cdot W^{uv} \\ \beta \cdot W^{vu} & W^{vv} \end{bmatrix}, \quad \beta = (1-\lambda_\gamma)/\lambda_\gamma$$

この問題を解き，さらに近似すると次のようになる．

$$F^* = (1-\mu_\alpha)(I - \mu_\alpha S)^{-1} F^0 \approx (1-\mu_\alpha)(I - \mu_\alpha \hat{S})^{-1} \hat{F}^0 \quad (7)$$

ここで，$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ であり，I は単位行列，$\mu_\alpha \in [0,1]$ はパラメータである．$F^0 = (f_1^0, \ldots, f_{m+n}^0)^T$ は初期適合度ベクトルを表す．$\hat{F}^0$ 及び $\hat{S}$ はそれぞれ F^0 及び S に対応するが，適合度推定のために構成する部分 2 部グラフから求める．部分 2 部グラフの構成方法の詳細は **3.1.1** で述べる．

3. 問題設定

本論文では比較的長いテキストデータが対応づけられたアイテムを対象とし，閲覧履歴や購入履歴などのユーザ行動履歴に基づいた情報推薦に着目する．以下では，例として Web ページの推薦を想定して議論する．このような情報推薦問題に取り組むにあたり，どのような情報が推薦されるべきかを考察する上で，次のような仮定が成り立つと考えられる．

1. ユーザは過去に閲覧した Web ページの内容とより類似している内容の Web ページを好む．
2. ユーザは他の行動が類似するユーザがより多く閲覧した Web ページを好む．

これらの仮定に基づくと，ある特定のユーザに着目したとき，そのユーザは，自身の Web ページの過去の閲覧履歴，及び，他の類似するユーザのページの閲覧履歴を考慮して推薦されたページを好むと考えられる．そこで，以上の 2 つの閲覧履歴を同時に考慮できるアルゴリズムとして，**2.** で述べた Co-HITS アルゴリズムを導入する．ユーザの Web ページに対する閲覧履歴が与えられるとき，それからユーザノードと Web ページノードからなる 2 部グラフを構成することができる．そして，所与のユーザのプロファイルと Web ページの内容情報から算出される適合度

を，グラフの構造に基づいて更新し，そうして得られた適合度によって Web ページをランク付けすれば上位 N 推薦が実現する．Co-HITS アルゴリズムを情報推薦に適用するにあたり，いくつかの変更点を加えた．以下ではその変更点について述べる．

3.1 Co-HITS アルゴリズムの情報推薦問題への適用

前述の目的のもと，ユーザノード集合 U とアイテム（例えば Web ページ）ノード集合 V からなる 2 部グラフを考え，所与のユーザに対する各アイテムの適合度を推定し，それらに基づいて推薦を行う．このために，Co-HITS アルゴリズムにおける部分行列の構成と初期適合度の計算に関して修正を適用したので，それぞれの詳細を以下に述べる．

3.1.1 部分行列の構成方法

本項では式 (7) における部分行列 $\hat{S}$ の構成方法について述べる．ある基準に基づいて元の 2 部グラフ G から部分グラフ $\hat{G}$ を抽出し，G から行列 S を求める代わりに，部分グラフ $\hat{G}$ から行列 $\hat{S}$ を求める．以下に，部分グラフ $\hat{G}$ の抽出手順を示す．

1. ノードの内容情報による初期適合度を元にノード集合 U，V それぞれについて上位 N_0 件のノードを取り出し，シード集合 $\hat{U} = U_L$，$\hat{V} = V_L$ とする．本論文では $N_0 = 10$ とした．
2. ノード集合 $\hat{V}$ に，$\hat{U}$ の各ノードと隣接するノードを追加し，更新する．
3. ノード集合 $\hat{U}$ についても同様に $\hat{V}$ の各ノードと隣接するノードを追加し，更新する．
4. 所定のサイズになるまで，上記の 2，3 を繰り返す．

なお，初期適合度の計算方法については次項で述べる．以上の手順により部分グラフを抽出した後，その部分グラフに対応する隣接行列を構成する．Deng ら [5] はクエリ推薦問題において，遷移行列の各要素である遷移確率を求める際，クリックスルー頻度（ある検索クエリの元で各 Web ページを閲覧した回数）に基づいて遷移確率を与えている．しかしながら，本論文で狙いとする情報推薦問題では，同一ユーザが同一アイテム（例えば Web ページ）を複数回にわたって閲覧するということが一般的でない状況も考えられるため，各ユーザが閲覧した各アイテムに対して一様な遷移確率を与えた．

3.1.2 初期適合度の計算方法

各ノードの初期適合度を求める方法として，Deng ら [5] は各ノードの内容情報を文書とみなしてユニグラム言語モデルで表現し，そのときのクエリ尤度 [7, 8] を初期適合度としている．これによってクエリと対象となる文書の類似度が高いほど 1 に，低いほど 0 に近い初期適合度が与えられている．本論文では問題設定の違いからこれを用いることは適していないと考えた．なぜなら，Deng らの問題設定では，単語数語から成るクエリなどの短いテキスト表現を対象としていたが，本論文では長いテキストデータ（例えば Web ページの内容，とくに後述の実験に用いたデータセットでは Web ニュースのヘッドラインまたは記事本文）が対応づけられたアイテムの推薦を目的としているため，その尤度を用いる場合に初期適合度の値が非常に小さくなってしまい有効でないからである．従って，本論文では一般的に広く用いられる分布間距離の一つである **Hellinger** 距離を用いることとした．2 つの n 次元ベクトル $\mathbf{p}, \mathbf{q}$ について，Hellinger 距離は次式で与えられる．

$$D_{HL}(\mathbf{q}, \mathbf{p}) = \sqrt{\frac{\sum_i^n (\sqrt{q_i} - \sqrt{p_i})^2}{2}} \tag{8}$$

なお，前処理として，各アイテムノードに対応づけられたテキストデータ（例えば Web ページ，とくに後述するデータセットの場合は Web ニュースのヘッドラインまたは記事本文）について形態素解析[1]を行った後，名詞のみに着目してユニグラム言語モデルを推定し，それを Hellinger 距離の計算に用いる．また，ユーザノードについては，各ユーザが過去に閲覧したアイテム（例えば Web ページ）のユニグラム言語モデルを上記と同様に推定した後，ユーザごとの期待値を求めて，それを Hellinger 距離の計算に用いる．Hellinger 距離は分布同士の類似度が高いほど値が 0 に近い値をとり，低いほど 1 に近い値をとる．つまり $0 \leq D_{HL} \leq 1$ の範囲をとるため，$1 - D_{HL}$ を初期適合度として用いた．

以上から，本論文において提案する情報推薦のための Co-HITS アルゴリズムは次の手順に従う．

1. ユーザノード集合 U とアイテムノード集合 V からなる 2 部グラフ $G = (U \cup V, E)$ を構成する．
2. 所与のユーザと各ノードの間の Hellinger 距離に基づいて初期適合度を計算し，U，V それぞれについて上位 N_0 件をシード集合 U_L，V_L として抽出する．$N_0 = 10$ とした．
3. **3.1.1** で述べた方法により，シード集合を用いて部分 2 部グラフ $\hat{G} = (\hat{U} \cup \hat{V}, \hat{E})$ を構成する．
4. $\hat{G}$ から初期適合度ベクトル F^0 及び行列 $\hat{S}$ を得る．
5. 式 (7) を解き，最終的な適合度ベクトル $\hat{F}^*$ を得る．
6. 適合度に基づいたアイテムのランク付けリストを出力する．

4. 実験

本論文では，Yahoo!ニュース閲覧履歴データを用いて，上位 N 推薦に基づいた実験的評価を行う．

4.1 データセット

2 部グラフを構成するためのデータセットとして Yahoo!ニュース閲覧履歴データを使用する．このデータはネットレイティングス社によって提供された 2010 年 6 月分の Web ページ閲覧履歴データ（ユーザ数 36218，ニュース数 1731）と国立情報学研究所によって収集された同時期の「Yahoo!ニュース」[2]の各ニュース記事のテキストデータを抽出したデータを組み合わせて構成されたものである．

4.2 実験設定

3. で述べた手法の有効性を確かめるため，評価実験を行う．その際，**1.** で述べたデータスパースネス問題やその極端な場合で

[1] 後述の実験では形態素解析に Igo を用いた（http://igo.sourceforge.jp/）．
[2] http://headlines.yahoo.co.jp

あるコールドスタート問題を想定し，以下の実験設定を行う．

実験では Yahoo!ニュース閲覧履歴データから無作為に 50 名のユーザを選択して用いる．ただし，ニュースの閲覧数が 50 件未満のユーザは除外する．これらのユーザについての閲覧履歴をもとに，閲覧履歴データから 1 割を抽出して評価のための正解データとし，これとは別途に 1 割を抽出してパラメータ設定のための開発データとする．残った 8 割の閲覧履歴（以下，訓練データ候補）のうち，訓練データとして実験に使用する割合を変えて実験を行う．これによって閲覧履歴が少ないユーザを想定することができ，本論文で述べた手法が前述のデータスパースネス問題にどの程度対応可能であるかを確認することができる．使用する割合については十分な閲覧履歴が利用可能であることを想定した 100%，データスパースネス問題が生じることを想定した 5% の 2 パターンについて実験を行う．後者の場合，各ユーザのニュース閲覧数が少なくとも 50 件であるため，閲覧数が最も少ない状況では閲覧数が $50 \times 0.8 \times 0.05 = 2$ 件となる．

4.2.1 パラメータ推定

Co-HITS をデータスパースネス問題が発生する状況に適用するため，Co-HITS の式 (7) の 2 つのパラメータ $\mu_\alpha, \lambda_\gamma$ の推定を行う．パラメータ推定には前項で述べた開発データを用い，この開発データに関する推薦結果を利用する．推薦処理に用いる閲覧データは，データスパースネス問題を考慮して訓練データ候補の 5% とする．まず，μ_α の推定を行う．λ_γ の値を 1 に固定し，μ_α の値を変えて実験を行う．実験結果の評価は P@5 に基づいて最適化する．これによって決定された μ_α を用いて同様に λ_γ の推定も実行する．

4.2.2 ベースライン手法

Web ページの内容のみを考慮したアルゴリズム，及び，2 部グラフの構造情報のみを用いたアルゴリズムと比較して Co-HITS アルゴリズムの有効性を評価する．以下に各々の詳細を述べる．

1. コンテンツベースフィルタリング
   Co-HITS アルゴリズムにおける初期適合度計算のみを行い，それによってアイテムノードのランク付けを行う．このとき，初期適合度の計算は Web ページの内容情報のみに基づいて算出されるものであるため，コンテンツベースフィルタリングの一種と見なすことができる．
2. 修正 HITS
   式 (7) においてパラメータを $\mu_\alpha = 1, \lambda_\gamma = 0.5$ とすることで，Co-HITS アルゴリズムにおける初期適合度による制約を無視しつつ，グラフ構造による大域的一貫性を考慮した各ノードの適合度計算が可能となる．これはグラフ構造によるノードのランク付けを行う HITS アルゴリズム [6] に相当し，協調フィルタリングの一種と見なすことができる．

4.2.3 評価指標

実験の評価指標には次の 3 つの指標を用いることとする．

1. N 位精度（Precision at top-N: P@N）
   N 位精度は上位 N 件のうち正解アイテムが含まれる比率として定義される．本実験では推薦結果上位に注目し，P@5, P@10 によって評価を行った．
   $$P@N = \frac{\text{上位 } N \text{ 件に含まれる正解データ数}}{N}$$
   最終的には上式による評価値を全ユーザに渡って平均する．
2. 平均精度（Average precision: AP）
   正解データ数が R 件のユーザに関する平均精度は次式で与えられる．
   $$AP = \frac{1}{R}\sum_{i=1}^{R} r_i \frac{i \text{ 番目までに含まれる正解データ数}}{i}$$
   ここで，r_i は i 番目の推薦結果が正解なら 1，そうでなければ 0 となるような関数を表す．最終的には上式による評価値について全ユーザに渡って平均をとる．
3. 逆順位（Reciprocal Rank: RR）
   逆順位は最も上位にランク付けされた正解アイテムの順位の逆数として定義される．どの程度上位に推薦されるべきアイテムがランク付けされたかによって評価を行う．
   $$RR = \frac{1}{\text{最上位にランク付けされた正解データの順位}}$$
   最終的には上式による評価値を全ユーザに渡って平均する．

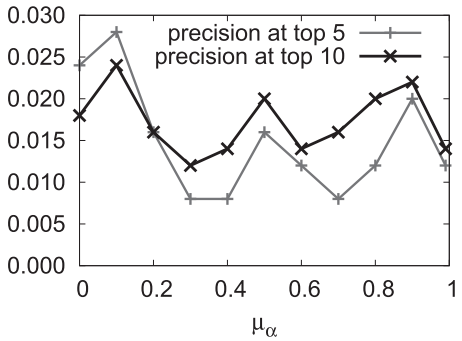

図 1 訓練データ候補を 5% 使用した場合の μ_α の影響
Fig. 1 Effect of μ_α with 5% of candidate training data.

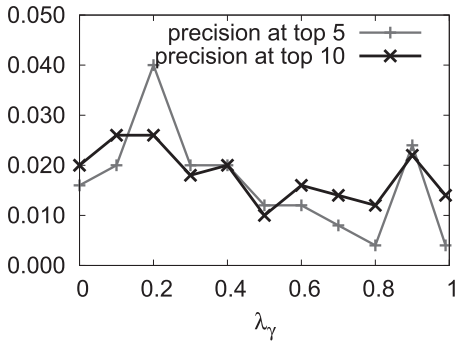

図 2 訓練データ候補を 5% 使用した場合の λ_γ の影響
Fig. 2 Effect of λ_γ with 5% of candidate training data.

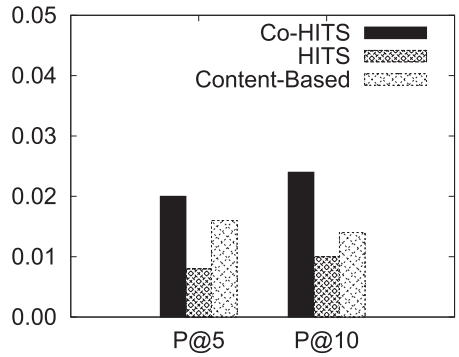

図 3　訓練データ候補の 5% 使用した場合の P@N による評価結果
Fig. 3　Evaluation results in terms of P@N with 5% of candidate training data.

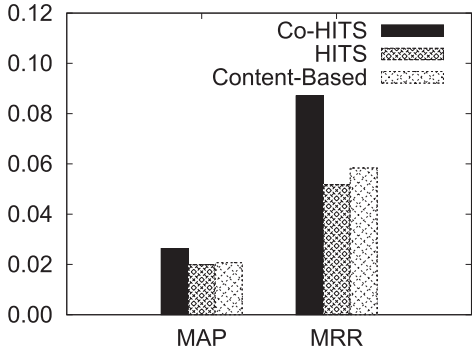

図 4　訓練データ候補の 5% 使用した場合の AP 及び RR による評価結果
Fig. 4　Evaluation results in terms of AP and RR with 5% of candidate training data.

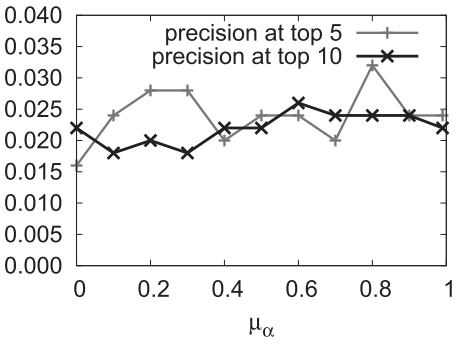

図 5　訓練データ候補を 100% 使用した場合の μ_α の影響
Fig. 5　Effect of μ_α with 100% of candidate training data.

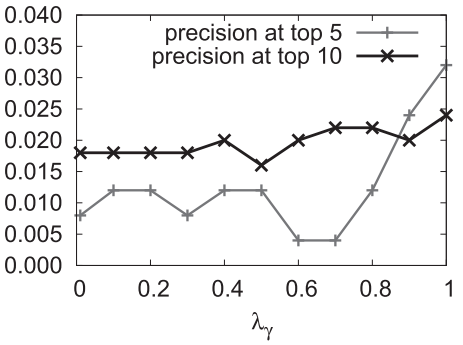

図 6　訓練データ候補を 100% 使用した場合の λ_γ の影響
Fig. 6　Effect of λ_γ with 100% of candidate training data.

4.3　実験結果

4.3.1　パラメータ推定（訓練データ候補の 5%）

訓練データ候補の 5% を使用した場合のパラメータ推定の結果について述べる．図 1，図 2 に 2 つのパラメータ $\mu_\alpha, \lambda_\gamma$ の値を変化させて行った実験の評価結果を示す．P@5 と P@10 とで類似した振る舞いを見せたが，より推薦結果の上位に着目し P@5 の結果を重視して $\mu_\alpha = 0.1$，$\lambda_\gamma = 0.2$ に決定した．

4.3.2　評価結果（訓練データ候補の 5%）

パラメータ推定によって得られた $\mu_\alpha = 0.1, \lambda_\gamma = 0.2$ を使用した Co-HITS アルゴリズムとコンテンツベースフィルタリング，及び，HITS によって得られた評価結果を図 3，図 4 に示す．これらの図から，それぞれの評価指標に関して Co-HITS がベースラインよりも良い性能を示すことがわかる．P@5 では HITS，コンテンツベースからそれぞれ 150%，25% の改善となり，P@10 ではそれぞれ 140%，71.4% の改善となった．また，AP と RR については各比較手法から 27.2% から 68.5% の改善となった．

なお，これらの改善の度合いは次式による．

$$\text{改善率} = \frac{\text{着目する手法の評価値 - ベースラインの評価値}}{\text{ベースラインの評価値}}$$

なお，AP に関して，Co-HITS は HITS とコンテンツベースフィルタリングのいずれに対しても，Wilcoxon 符号付順位検定及び対応のある t 検定の両方で有意水準 0.05 の有意差が認められた．他の評価指標では有意差は認められなかったが，それらの評価指標は十分に大きな評価値サンプル数でなければ有意差が表れにくいことが知られている [9]．以上の結果から，データスパースネス問題が起こる条件下において提案手法が有効であると言える．

4.3.3　パラメータ推定（訓練データ候補の 100%）

訓練データ候補の 5% を用いた実験と同じ要領で，訓練データ候補をすべて使用する条件下でパラメータ推定を実行した．このときの μ_α，λ_γ について，それぞれ図 5，図 6 に示す．P@5 と P@10 とでやや異なる振る舞いが見られるが，より推薦上位に着目した P@5 を重視し，$\mu_\alpha = 0.8$，$\lambda_\gamma = 1.0$ に決定した．

4.3.4　評価結果（訓練データ候補の 100%）

訓練データ候補の 5% を用いた実験と同じ要領で得られた評価結果を図 7 と図 8 に示す．これらから，閲覧履歴が十分に与えられた場合は AP に関して Co-HITS がベースラインよりもやや優

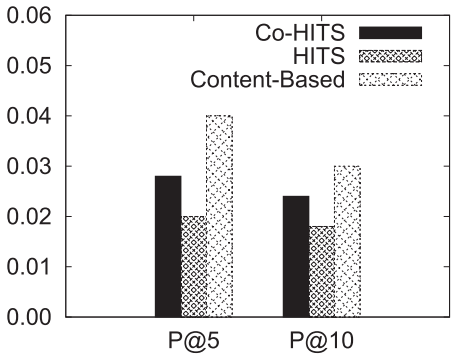

図 7　訓練データ候補を 100% 使用した場合の P@N による評価結果
Fig. 7　Evaluation results in terms of P@N with 100% of candidate training data.

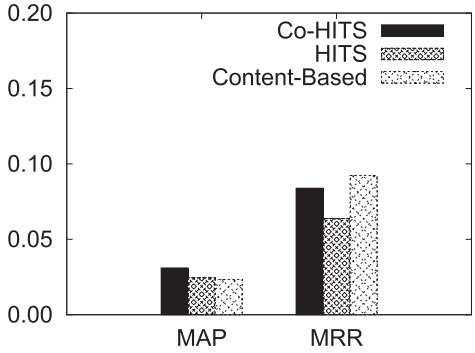

図 8　訓練データ候補を 100% 使用した場合の AP 及び RR による実験結果
Fig. 8　Evaluation results in terms of AP and RR with 100% of candidate training data.

れているものの，それ以外の評価指標に関してはコンテンツベースが最も有効であった．以上のことから，データスパースネス問題を意図的に排除した状況下ではコンテンツベースが有効であると言える．

5.　おわりに

本論文では比較的長いテキストデータが対応づけられたアイテムを対象とし，その内容情報と閲覧履歴や購入履歴などのユーザ行動履歴を，正則化付き相互強化の枠組みを用いて組み合わせた情報推薦手法を提案した．本手法は Co-HITS アルゴリズムに基づくものであり，クエリ推薦を目的として設計された当該アルゴリズムを修正し，前述のような情報推薦の問題に適用したものである．実験によって，このアルゴリズムが情報推薦問題でしばしば問題となるデータスパースネス問題が生じる条件下でよい推薦結果をもたらし，安定した推薦を実現できることを示した．他の推薦手法との比較や考察については今後の課題である．

【謝辞】

本研究の一部は，科学研究費補助金（23300039, 22240007）の援助による．実験データを提供して頂いた国立情報学研究所の韓浩氏，中渡瀬秀一氏，大山敬三氏に感謝する．

【文献】

[1] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*, Cambridge University Press, 2010.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, *GroupLens: an open architecture for collaborative filtering of netnews*, Proc. of the 1994 ACM conference on Computer supported cooperative work, pp.175.186, 1994.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, *Item-based collaborative filtering recommendation algorithms*, Proc. of the 10th international conference on World Wide Web, pp.285.295, 2001.

[4] A.I. Schein, A. Popescul, L.H. Ungar, and D.M. Pennock, *Methods and metrics for cold-start recommendations*, Proc. of the 25th annual international ACM SIGIR conference, pp.253.260, 2002.

[5] H. Deng, M.R. Lyu, and I. King, *A generalized Co-HITS algorithm and its application to bipartite graphs*, Proc. of the 15th ACM SIGKDD international conference, pp.239.248, 2009.

[6] J.M. Kleinberg, *Authoritative sources in a hyperlinked environment*, Proc. of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, p.668, 1998.

[7] D. Hiemstra, *A linguistically motivated probabilistic model of information retrieval*, Proc. of the 2nd European conference on Research and advanced technology for digital libraries, pp.569.584, 1998.

[8] J.M. Ponte and W.B. Croft, *A language modeling approach to information retrieval*, Proc. of the 21st annual international ACM SIGIR conference, pp.275.281, 1998.

[9] C. Buckley and E.M. Voorhees, *Evaluating evaluation measure stability*, Proc. of the 23rd annual international ACM SIGIR conference, pp.33.40, 2000.

内藤 慎也 Shinya NAITO
西日本高速道路株式会社勤務．平成 23 年神戸大学工学部情報知能工学科卒業．平成 25 年同大学大学院システム情報学研究科情報科学専攻博士前期課程修了．

江口 浩二 Koji EGUCHI
神戸大学大学院システム情報学研究科准教授．博士（工学）．情報検索，統計的機械学習，データマイニングの研究に従事．