Vol.2011-DBS-153 No.20
2011/11/3

正則化付きリンク構造解析を用いたコールドスタート推薦

内藤慎也[†1]　江口浩二[†1]

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．とりわけ本研究では比較的長いテキストを持つアイテムを対象とした情報推薦について，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 部グラフの正則化付きリンク構造解析によって実現することを目指す．本手法は，グラフ構造とともにユーザプロファイルとアイテムのコンテンツを相互強化の枠組みで統合するものであり，情報推薦で現在主流となっている手法にはなかったものである．実験では，Web 閲覧履歴データを使用し，Web ニュースを対象アイテムとした．ユーザ毎にヘルドアウトした一部の被閲覧アイテム集合からなるテストセットを使用し，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した．実験の結果，情報推薦においてしばしば問題となるコールドスタート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と比較してより有効であることを示す．

Cold-start Recommendations using Regularized Link Analysis

Shinya Naito[†1] and Koji Eguchi[†1]

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 with long text, and aim at achieving this by proposing a method of link analysis of a user-item bipartite graph in a regularization framework based on Co-HITS algorithm. This method can integrate, via mutual reinforcement, the graph structure and the content of both user profiles and items. It has never been seen in the mainstream of conventional recommendation techniques. In our experiments, we used the data of Web browsing history, assuming Web news articles as target items. We evaluated the list of top-N items recommended based on the browsing history, using a test set that consists of a part of browsed items for each user. We demonstrate through the experiments that the proposed method outperformed several baseline methods in a situation where the cold start problem occurs, which often becomes a serious problem in recommendations.

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が過去にどのような行動をとったか (どのような商品を購入したか，どんなニュースを閲覧したか)，あるいは同じような行動をとった他のユーザがその前後にどのような行動をとっているかなどによって推薦を行うものである．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**協調フィルタリング**，**コンテンツベースフィルタリング**の 2 つがある[1),2)]．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，比較的新しいユーザや比較的新しいアイテムが出現するとそれらについての情報が少なく十分な推薦ができないという問題が発生する．これは，推薦を行うシステムが上記のような手法を用いて推薦したい対象がこれまでにどのような行動をとっているか，または推薦したい対象に類似しているもの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．これを**コールドスタート問題**と呼ぶ．我々はこのコールドスタート問題に注目し，これを解決するために Co-HITS アルゴリズム[3)] 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，そのまま情報推薦に適用するのは適切でないため，種々の観点から改良を行う．そして，実験では情報推薦や情報検索において用いられているコンテンツベースフィルタリング，HITS に相当したアルゴリズムと比較し，コールドスタート推薦が起こる状況下において提案手法がより優れていることを示す．

2. Co-HITS アルゴリズム

本節では本研究で提案する Co-HITS アルゴリズムの概要について説明する．まず Co-HITS を適用するために必要な 2 部グラフについて簡潔に説明した後，Co-HITS の定式化について述べる．なお，Deng らの研究では，クエリと Web ページによる 2 部グラフを考え，第 3 章我々の提案手法では

2.1 2 部グラフとランダムウォーク

2 部グラフは 2 つのデータの集合間の関係を表すために広く用いられてきた．2 部グラフ $G = (U \cup V, E)$ を考えるとき，各ノードは独立した 2 つの集合 U，V に分けられる．そして，

†1 神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

辺は片方の集合のノードからもう片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，それぞれ m 個と n 個のノードをもつ2つの集合 $U = \{u_1, u_2, \ldots, u_m\}$, $V = \{v_1, v_2, \ldots, v_n\}$ 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 が存在し，さらにそれらのノードを結ぶ辺があるとき，u_i と v_j の間の遷移確率を w_{ij}^{uv} および 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値が 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 となり，$\Sigma_{j \in V} w_{ij}^{uv} = 1$ 及び $\Sigma_{i \in U} w_{ji}^{vu} = 1$ が得られる．

さらに，これらの遷移確率から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も，隠れ遷移確率 w_{ij}^{uu} および w_{ij}^{vv} を考えることができる．u_i から u_j への隠れ遷移確率 w_{ij}^{uu} は次式で求めることができる．

$$w_{ij}^{uu} = \sum_{k \in V} w_{ik}^{uv} w_{kj}^{vu}, \tag{1}$$

そして，

$$\sum_{j \in U} w_{ij}^{uu} = \sum_{j \in U} \sum_{k \in V} w_{ik}^{uv} w_{kj}^{vu} = \sum_{k \in V} \left(w_{ik}^{uv} \sum_{j \in U} w_{kj}^{vu} \right) = \sum_{k \in V} w_{ik}^{uv} = 1 \tag{2}$$

逆の v_i から v_j への隠れ遷移確率 w_{ij}^{vv} についても同様に求めることができる．

これらの遷移確率を用いてランダムウォークを 2 部グラフ上で考えることができるので，例えば U から V への遷移について遷移行列 $W^{uv} \in \mathbf{R}^{m \times n}$ を作成する．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} であり，逆の V から U の遷移行列 W^{vu} も同様に示すことができる．さらに，$W^{uu} \in \mathbf{R}^{m \times m}$ と $W^{vv} \in \mathbf{R}^{n \times n}$ をそれぞれ U と V それぞれの内部における隠れ遷移行列を示すために使用する．

2.2 Co-HITS アルゴリズム

Co-HITS アルゴリズムは Deng らによって提案されたアルゴリズムである[3]．Deng らはクエリ推薦を目的としてクエリと Web ページから 2 部グラフを構成し，この 2 部グラフを用いてクエリ間の類似性を求めることができるとした．以下では，Deng らの論文[3] に従って，Co-HITS の概要を述べる．

2 部グラフの各ノードについて，他ノードへの遷移確率によるランダムウォークを考えることができる．この遷移確率を用いて反復操作を繰り返し行うことによってグラフ内でのスコアの伝達が可能となる．さらに各ノードはノードの内容情報を持ち，クエリ (3 章で述べる提案手法ではユーザープロファイルに注目する) が与えられたときにテキスト間の適合度を計算する関数を用いることによって適合スコアを計算することができる．Co-HITS アルゴリズムはこれらを組み合わせることにより，ノードの内容情報及びグラフの構造の両方を考慮したクエリとの適合度計算を行うことを可能としている．Co-HITS アルゴリズムによる u_i のスコア x_i と v_k のスコア y_k は式 (3) , (4) のように定義できる．

$$x_i = (1 - \lambda_u) x_i^0 + \lambda_u \sum_{k \in V} w_{ki}^{vu} y_k, \tag{3}$$

$$y_k = (1 - \lambda_v) y_k^0 + \lambda_v \sum_{j \in U} w_{jk}^{uv} x_j, \tag{4}$$

ここで，$\lambda_u \in [0, 1]$ と $\lambda_v \in [0, 1]$ はノードの内容情報とグラフ構造の情報それぞれの重みを決定するパラメータ，x_i^0 と y_k^0 は u_i と v_k それぞれの初期スコアである．このモデルにおいて，初期スコアは標準化し $\Sigma_{i \in U} x_i^0 = 1$ と $\Sigma_{k \in V} y_k^0 = 1$ とする．これより反復操作の後，x_i の合計と y_k の合計はそれぞれ 1 になる．

前述の反復操作について，各ノードは近傍から伝搬されるスコアを受け取り，かつノードの初期スコアを保持している．反復操作が繰り返されると，各ノードの初期スコア及びその近傍によって決定されるあるスコアに収束する．この反復過程において，Co-HITS アルゴリズムでは正則化を行う．ここでは，初期ランキングスコアに基づく正則化項を導入することにより，グラフ全体の適合スコアについての正則化を行うことができる[4]．以下でその詳細について説明する．2 部グラフのグループ U について，コスト関数 R_1 は次のように定義できる．

$$R_1 = \frac{1}{2} \sum_{i,j \in U} w_{ij}^{uu} \left\| \frac{x_i}{\sqrt{d_{ii}}} - \frac{x_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i \in U} \| x_i - x_i^0 \|^2 \tag{5}$$

ここで $\mu > 0$ は正則化パラメータであり，D は標準化のための対角行列で各要素 d_{ii} が $d_{ii} = \Sigma_j w_{ij}$ である．コスト関数の第 1 項目はグラフ全体について改良されたランキングスコアの大域的一貫性を定義し，式 (3) のスコア伝達を定義する第 2 項目に相当する．式 (5) の 2 項目は初期ランキングスコアによる制約を定義し，式 (3) の初期スコアを保持する第 1 項目に相当する．たがい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 は次のようにあらわされる．

$$R_2 = \frac{1}{2} \sum_{i,j \in V} w_{ij}^{vv} \left\| \frac{y_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i \in V} \| y_i - y_i^0 \|^2, \tag{6}$$

ここでの正則化の背景には，類似したノードが最も類似した適合スコアをもつであろうという大域的一貫性の考えがある．R_1 と R_2 では U と V それぞれのグループの内部隠れリンクに基づいた整合性が定められていたが，ここで U と V 間の直接リンクがスコア伝搬と相互強化においてより大きな影響力をもつと考えられ，U と V の直接的な関係を考慮するために新しいコスト関数 R_3 提案された．

$$R_3 = \frac{1}{2}\sum_{i\in U, j\in V} w_{ij}^{uv}\left\|\frac{x_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}}\right\|^2 + \frac{1}{2}\sum_{j\in V, i\in U} w_{ji}^{vu}\left\|\frac{y_j}{\sqrt{d_{jj}}} - \frac{x_i}{\sqrt{d_{ii}}}\right\|^2 \quad (7)$$

R_3 の背景には，2 つの集合間の平滑化の制約がある．これは，強く結びついた U と V のノードの適合スコアの大きな差にペナルティーを課すというものであり，U と V の双方から関連付けられたコスト関数 R は次のように定義されている．

$$R = \lambda_\gamma(R_1 + \alpha R_2) + (1-\lambda_\gamma)R_3 \quad (8)$$

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，正則化付き Co-HITS の実現が可能となる．Deng らの論文[3] では，$\alpha = 1$ とし，パラメータ $\lambda_\gamma = 0.5$ が最適であることが示されている．

また，これらを一般化した最適化問題 $\min_F(R)$ は次のように書きなおすことができる．

$$\begin{aligned} \min_F \quad & \frac{1}{2}\sum_{i,j=1}^{m+n} w_{ij}\left\|\frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}}\right\|^2 + \mu\sum_{i=1}^{m+n} \| f_i - f_i^0 \|^2 \\ s.t. \quad & W = \begin{bmatrix} W^{uu} & \beta\cdot W^{uv} \\ \beta\cdot W^{vu} & W^{vv} \end{bmatrix} \\ & F = \begin{bmatrix} X \\ Y \end{bmatrix} \\ & \beta = (1-\lambda_\gamma)/\lambda_\gamma \end{aligned} \quad (9)$$

ここで，X と Y はそれぞれのスコアベクトルである．この問題を解く[5)–7)] と，次式を導くことができる．

$$\begin{aligned} & F^* = \mu_\beta(I - \mu_\alpha S)^{-1}F^0 \\ & \mu_\alpha = \frac{1}{1+\mu}, \mu_\beta = \frac{\mu}{1+\mu} \end{aligned} \quad (10)$$

ここで，I は単位行列，$S = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$，$\mu_\alpha$ の範囲は 0 から 1，$\mu_\alpha + \mu_\beta = 1$ である．この式により，反復計算を行わずに直接に適合スコア F^* を導くことができる．

ここで，記憶領域と逆行列の計算速度のバランスをとるための方法として，式 10 の行列 S を部分行列 $\hat{S}$ で置き換えることとする．この $\hat{S}$ は初期適合スコアで上位 n 件にランクしたノード $\hat{F^0}$ を用いて導かれる．部分行列の導出法は後述する．ここで上位 n 件より下位のスコアを 0 に近いとみなすと，式（10）は次の近似式と等価とすることができる．

$$\hat{F^*} = (I - \mu_\alpha\hat{S})^{-1}\hat{F^0} \quad (11)$$

ここでランキングに直接関係がないため，μ_β は省略されている．実際の計算にはこの式を用いることとする．

3. 問題設定

情報推薦問題に取り組むにあたり，本研究においてどのような情報が推薦されるべきであるかの定義を次のような仮定に基づいて考える．

(1) ある着目するユーザに関して，その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) ある着目するユーザに関して，そのユーザは他の行動が類似するユーザからよりたくさん閲覧されている Web ページを好む．

これらの仮定に基づくと，あるユーザが入力として与えられたときそのユーザは自身の Web ページの過去の閲覧履歴，さらに他のユーザのページの閲覧履歴を考慮し，その結果推薦されたページを好むと推測することができる．そこで，以上 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，Co-HITS アルゴリズムを導入する．複数の Web ページの内容情報及び複数のユーザの Web ページの閲覧履歴情報がデータセットとして与えられるとき，それらの情報から 2 部グラフを構成することができる．これらの情報から構成される 2 部グラフの形を**図 1** に示す．このアルゴリズムは 2 部グラフが構成できるデータに対して用いられるものであるため，図 1 で示したような 2 部グラフに対しても適用することが可能である．そして，そこから算出されるスコアに基づいてページをランキングすれば入力に対する推薦されるべきページを得ることができる．Co-HITS アルゴリズムを本研究に適用するにあたり，いくつかの変更点を加えた．その変更点について述べてゆく．

3.1 Co-HITS アルゴリズムの情報推薦問題への適用

本節では，これまでに述べた Co-HITS アルゴリズムを，Web ページの推薦問題に適用する方法について述べる．

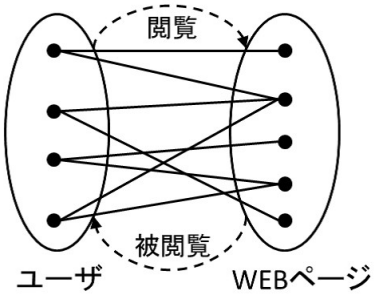

図 1 データセットを適用した 2 部グラフ例
Fig. 1 An example of bipartite graph from a dataset

3.1.1 部分行列の導出

本項では式（11）に用いるための部分行列 $\hat{S}$ の導出について述べる．部分行列は元の 2 部グラフから部分グラフを取り出し，その部分グラフについての隣接行列を取り出すことで作成することができる．以下に，部分グラフ $\hat{G}$ の作成手順を示す．

(1) ノードの内容情報による初期ランキングスコアを元に各集合についてトップ 10 件のノードを取り出し，シードセット $\hat{U} = U_L$，$\hat{V} = V_L$ としてセットする．
(2) 片側のセット $\hat{V}$ について，$\hat{U}$ のノードと辺がつながっているノードを追加し，更新する．
(3) もう片方のセット $\hat{U}$ についても同様に $\hat{V}$ のノードと辺がつながっているノードを追加し，更新する．
(4) 欲しいサイズになるまで，上記 2,3 を繰り返す．
(5) 欲しいサイズになったら終了する．

以上の仮定から部分グラフを取り出した後，取り出した部分グラフについて隣接行列を取り出すが，このとき元論文では各ノードについて広く用いられている k-近傍法 (k-NN) を適用している．k-近傍法は各ノードについて，遷移確率の値を元に k 番目までの近傍に接続しているとみなす手法である．これにより得られる部分行列はたいていとても疎となるので，逆行列の計算にかかる時間を減少させることが可能である．ここで，Deng ら[3] は隣接行列の各要素である遷移確率を求める際，使用するデータに含まれるクリック回数 (検索クエリを元にどの Web ページを何回閲覧したか) を使用して遷移確率に重みをつけている．しかし本研究で狙いとする Web ページの推薦問題では，同一ユーザが同一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と考えられる．以上のことから，k-近傍による隣接行列値での選択が困難であるとし，本研究では全値を用いることとしている．

3.1.2 初期スコアの計算方法

各ノードの初期スコアを求める方法として，元論文では各ノードの内容情報を文書とみなして統計的言語モデルを適用し，クエリ尤度による値を使用している．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 に近い初期スコアが与えられている．本研究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．なぜならば元の問題設定では，単語数語から成るクエリなどの短い文書を対象としていたが，本研究では長い文書 (Web ページの内容，後述の実験に用いたデータセットであれば Web ニュース本文) を対象としている．そのため，クエリ尤度では初期スコアの値が非常に小さくなってしまい有効な値を得ることができないからである．よって，本稿では一般的によく用いられる情報量である **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 距離は式

$$D_{HL}(\mathbf{q}, \mathbf{p}) = \sqrt{\frac{\sum_i^n (\sqrt{q_i} - \sqrt{p_i})^2}{2}} \tag{12}$$

で与えられる．なお，各ページの Hellinger 距離は，前処理として各 Web ページの本文 (用いたデータセットの場合では yahoo ニュースの記事本文) について形態素解析を行った後，名詞のみに着目して各ニュースの単語分布を求め，それらの分布を上式のベクトルとみなしそれらについて計算を行うこととした．また，ユーザの閲覧履歴との Hellinger 距離については，ユーザの閲覧した各ページの単語分布を上記同様求めた後，ユーザごとにその期待値を求めることでユーザが持つ単語分布とし，Hellinger 距離の計算を行う．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，初期スコアの値を $1 - D_{HL}$ とした．これにより，類似度が高ければ 1 に近い値，低ければ 0 に近い値をとる適合スコアの初期値として扱うことができる．

以上から，本研究において提案する情報推薦のための Co-HITS アルゴリズムの手順を次のようにした．

(1) 入力として推薦対象ユーザ，2 部グラフを受け取る．
(2) Hellinger 距離に基づいて初期ランキングスコアを計算し，上位 10 件にランクづけられたものをシードセット U_L，V_L として抜き出す．
(3) 部分 2 部グラフ $\hat{G} = (\hat{U} \cup \hat{V}, \hat{E})$ へ拡張を行う．

表 1 Yahoo ニュース閲覧履歴データの概要
Table 1 Overview of Yahoo news browsing history data

	2010 年 6 月
ユーザ数	36218
ニュース数	1731

(4) 部分 2 部グラフから $\hat{S}$ を得て，初期スコアベクトル F^0 を得る．
(5) 式（11）を解き，最終スコア $\hat{F}^*$ を得る．
(6) 推薦対象ユーザに基づく Web ページの順位付け結果を出力する．

4. 実　　験

本節では，Yahoo!ニュース閲覧履歴データを用いて各評価手法による手法の比較を行う．

4.1 データセット

本研究では，2 部グラフを構成するデータセットとして yahoo ニュース閲覧履歴データを使用した．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と国立情報学研究所によって収集された実際の yahoo ニュースの各ニュースページ内容を抽出したデータを組み合わせて作成されたものである．本研究に用いたデータセットの詳細について**表 1** に示す． このデータセットは 2 つのファイルから成る．1 つ目は 6 月分の各 yahoo ニュースに独自 ID がふられそのニュース内容などとともに保存されているファイル，さらに 2 つ目はそのニュースそれぞれについてネットレイティングス社に登録され ID が割り振られたユーザが閲覧したかどうかを数値で表すファイルである．データの形式について**表 2**，**表 3** に示す．

4.2 データ前処理

実際に実験を行うにあたり，あらかじめ各ユーザの閲覧履歴を調査し，ニュースの閲覧数が 50 件未満のユーザは除去した．さらにそれぞれのユーザから 20%を取り出して正解データとすることとした．取り出した閲覧履歴は実際のユーザのデータから取り除き，仮想的にそのページを閲覧していないものとした上で正解データの予測を行う．

4.3 実 験 設 定

本研究の有効性を確かめるため，評価実験を行う．その際，以下の点について着目して評価を行う．

協調フィルタリングにおいて推薦を行う際，例えば一般的である相関係数法を用いると推薦対象と既存ノードで類似度の計算を行うが，この際に推薦対象の情報が少ない場合には類似度の計算が十分に行われないため十分な推薦が行えない．この問題は序論でも述べたようにコールドスタート問題と呼ばれる．一方で，コンテンツベースフィルタリングなどではこの問題は発生しない．本研究でもグラフ構造及び内容情報を同時に用いることができるため，コールドスタート問題を軽減できることが期待される．よって，コールドスタート問題を意識した問題設定とするために以下のような実験設定を行った．

実験では yahoo!閲覧履歴データセットを用い，使用するデータの中から無作為に 50 人のユーザを選択し，このユーザについての閲覧履歴をもとに，正解として閲覧履歴のデータを 2 割取り出す．そして残りの 8 割を閲覧履歴としてデータの予測を行う．このときに 8 割残した閲覧履歴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．これによって人工的に閲覧履歴が少ないユーザを作りだすことが可能であり，上述のコールドスタート問題について本研究がどの程度対応可能であるかを確認することが可能である．使用する割合についてはコールドスタート問題が発生しない十分な閲覧履歴の情報がある 100%，コールドスタート問題が発生する 5%の 2 パターンについて実験を行った．各ユーザは最低ニュース閲覧数が 50 件であるため，一番閲覧数が少なくなる状況を考えると閲覧数が 50

表 2 Yahoo ニュース記事データ形式
Table 2 Data form of yahoo news article data

記事 ID	記事第 1 段落	記事全文	トピック	メインカテゴリ	サブカテゴリ
article0001	article0001 の第 1 段落	article0001 の記事全文	芸能界	エンターテイメント	エンタメ総合
article0002	article0002 の第 1 段落	article0002 の記事全文	2010 年民主党代表選挙	国内	政治
...	...	...	...	...	...
articlexxxx	articlexxxx の第 1 段落	articlexxxx の記事全文	番組情報	エンターテイメント	エンタメ総合

表 3 Yahoo ニュース閲覧情報データ形式
Table 3 Data form of yahoo news browsing history

モニター ID	サンプル	article0001	article0002	...	articlexxxx
17458301	Work	2	0	...	1
13575468	Home	0	2	...	0
...	...	...	...	...	...
99999999	Home	1	1	...	1

Vol.2011-DBS-153 No.20
2011/11/3

件のユーザがいる場合の $50 \times 0.8 \times 0.05 = 2$ 件となる．この閲覧数 2 件は他の論文でもよく使用されているコールドスタート問題が発生する条件を満たしている[8),9)]．

実験で使用するにあたって実装を行った Co-HITS アルゴリズムにおいて，前述したように隣接行列を求めたのちに式（11）の行列計算を実行するが，本研究においては実装に java を用いたため java の疎及び密行列クラスである Universal Java Matrix Package(UJMP)[⋆1]を用いた．さらに，初期スコア導出におけるページ本文の形態素解析の実行については java のフリー形態素解析エンジン Igo[⋆2]を用いた．実験はまず Co-HITS のパラメータ推定を行った後に，評価実験として推定したパラメータを用いた提案手法と後述の 2 つの手法について実行した．

4.3.1 パラメータ推定

Co-HITS をコールドスタート推薦条件下用にカスタマイズするため，Co-HITS の式（11）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正解データのうちの半分を利用してこれを開発データとし，この開発データの推薦結果を利用する．利用する閲覧データはコールドスタート推薦を考慮した本実験と同様に，閲覧データの 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．

4.3.2 評 価 実 験

(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スコア計算部のみを実行し，このスコアによって各ノードのランク付けを行う．初期スコアの計算時に Hellinger 距離を用いており，これは Web ページの内容情報をもとに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と等価であるといえる．

(2) correspond to HITS
Co-HITS の式 11 において，パラメータを $\mu_\alpha = 1, \lambda_\gamma = 0.5$ とすることで初期スコアが無視され，Co-HITS アルゴリズムのグラフ構造による大域的一貫性を考慮したスコア付けが可能となる．これはグラフ構造によるスコア付けを行う HITS アルゴリズム[10)] に相当するものとなる．

Web ページの内容のみを考慮したアルゴリズム及び，Web ページの内容情報を用いず 2 部グラフの構造情報のみを用いたアルゴリズムを比較の対象とすることで Co-HITS アルゴリズムの有用性の評価を行う．

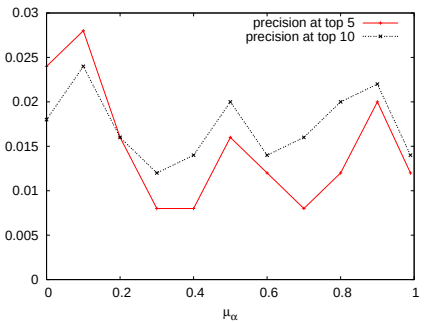

図 2 閲覧履歴を 5%使用した場合の μ_α の影響
Fig. 2 The effect of varying μ_α using 5% of browsing history

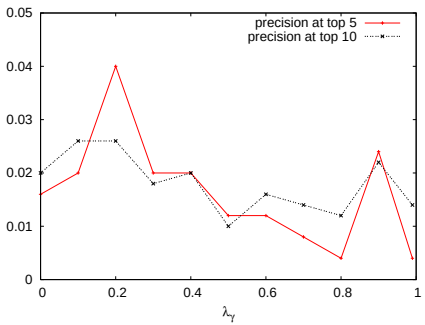

図 3 閲覧履歴を 5%使用した場合の λ_γ の影響
Fig. 3 The effect of varying λ_γ using 5% of browsing history

4.3.3 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) Precision at top-N($P@N$)
トップの N 件のうち，正解アイテムがどの程度含まれているのかの値をとったもの．情報検索の分野で用いられることが多い評価手法であり，正解アイテムがトップの結果にたくさん存在するほどその結果が有効であるという考えに基づく評価手法である．本実験では推薦結果上位に注目し，P@5，P@10 で評価を行った．

$$P@N = \frac{\text{上位 } N \text{ 件内にランキングされた正解データ数}}{N}$$

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]．

(2) Mean average precision(MAP)
適合率 (Average Precision:AP) の全評価ユーザの平均をとった値．正解データ数が n 件のユーザの適合率は次式で与えられる．

⋆1 http://www.ujmp.org/
⋆2 http://igo.sourceforge.jp/

⋆3 後述の Mean Average Precision や Mean Reciprocal Rank の呼称に倣うならば Mean Precision とすべきであるが，慣例に従って単に Precision と表記する．

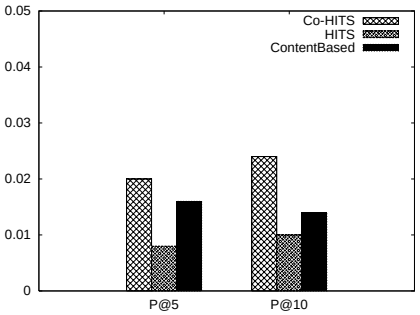

図 4 閲覧履歴を 5%使用した場合の P@N による実験結果
Fig. 4 Experimental results in terms of P@N using 5% of browsing history

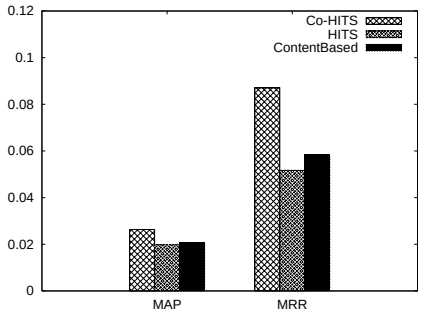

図 5 閲覧履歴を 5%使用した場合の MAP，MRR による実験結果
Fig. 5 Experimental results in terms of MAP and MRR using 5% of browsing history

$$AP = \frac{1}{\text{正解データ数}} \sum_{i=1}^{n} r_i \times \frac{\text{i 番目までに含まれる正解データ数}}{i}$$

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) Mean Reciprocal Rank(MRR)
最も上位にランクした正解アイテムの順位の逆数の値をとったもの．情報推薦においてできうる限りトップに近い順位に 1 件でも推薦されるべきアイテムが存在すれば，推薦結果としてより有意であるという考えによる評価手法である．どの程度トップに近いところに推薦されるべきアイテムが存在しているかによって評価を行う．

$$RR = \frac{1}{\text{ユーザの最も上位にランキングされた正解データの順位}}$$

P@N や MAP と同様，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．

4.4 実験結果

4.4.1 パラメータ推定 (5%)

ここでは 5%の閲覧履歴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 2**，**図 3** に 2 つのパラメータ $\mu_\alpha, \lambda_\gamma$ の値を変えながら実験を行った結果を示す．まず μ_α について図 2 に示す．P@5 と P@10 で類似した振る舞いを見せており，双方でよい値を示した $\mu_\alpha = 0.1$ とすることとした．この値を使用して，次に λ_γ の推定を行った．これを図 3 に示す．λ_γ も P@5，P@10 でほぼ類似した振る舞いを示したが，より推薦上位に着目し P@5 の結果を重視することとした．その結果，$\lambda_\gamma = 0.2$ とすることとした．

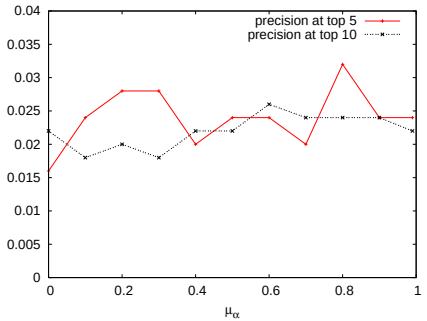

図 6 閲覧履歴を 100%使用した場合の μ_α の影響
Fig. 6 The effect of varying μ_α using 100% of browsing history

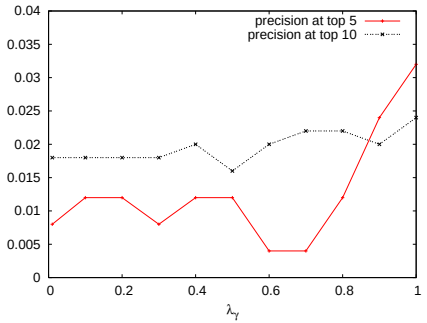

図 7 閲覧履歴を 100%使用した場合の λ_γ の影響
Fig. 7 The effect of varying λ_γ using 100% of browsing history

4.4.2 評価実験（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，HITS によって得られた結果を**図 4**，**図 5** に示す．グラフから，それぞれの評価指標について提案手法が比較手法よりも良い結果を示すことが示された．P@5 では HITS，コンテンツベースからそれぞれ 150%，25%の向上となり，P@10 ではそれぞれ 140%，71.4%の向上となった．また，MAP と MRR については各比較手法から 27.2%から 68.5%の向上となった．なお，これらのパーセントで表される改善の度合いについては，百分率で表現することのできる改善率による結果である．改善率は次式で示される．

$$\text{改善率} = \frac{\text{提案手法による評価値-比較手法による評価値}}{\text{比較手法による評価値}} \times 100\,[\%]$$

なお，MAP の評価指標について，提案手法は 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 符号付順位検定および対応のある t 検定により有意水準 0.5 で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，これらの評価指標は十分に多い評価値サンプル数でなければ有意差がでないことが知られている．以上の結果

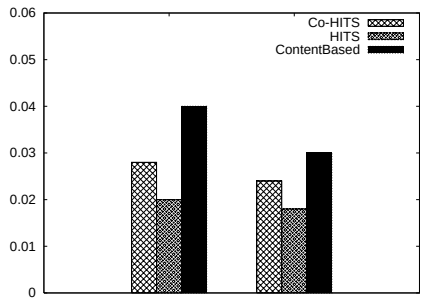

図 8 閲覧履歴を 100%使用した場合の P@N による実験結果
Fig. 8 Experimental results in terms of P@N using 100% of browsing history

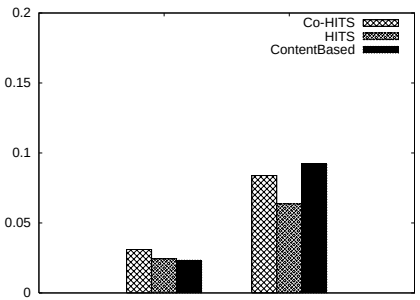

図 9 閲覧履歴を 100%使用した場合の MAP，MRR による実験結果
Fig. 9 Experimental results in terms of MAP and MRR using 100% of browsing history

から，コールドスタート問題が発生するような条件下において提案手法が効果的であることが示された．

4.4.3 パラメータ推定（100%）

5%のときと同様に，閲覧履歴をすべて使用する条件下のパラメータ推定を実行した．μ_α について，**図 6** に示す．P@5 と P@10 で少し異なる振る舞いを見せているが，より推薦上位に着目した P@5 を重視し，$\mu_\alpha = 0.8$ とすることとした．この値を使用して，λ_γ の推定を行った．**図 7** に示す．λ_γ も同様に P@5 の結果を重視し，$\lambda_\gamma = 1.0$ とすることとした．

4.4.4 評価実験（100%）

5%の場合と同様にして得られた結果を**図 8**，**図 9** に示す．グラフから，閲覧履歴をすべて使用する場合は MAP の評価指標では提案手法が優れており，それ以外の評価指標ではコンテンツベースが優れていた．以上の結果から，コールドスタートユーザ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ことがわかる．

5. おわりに

この論文では Web ページの内容情報とページの閲覧履歴情報に基づくグラフによる情報を，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るが，クエリ推薦を目的とした当該アルゴリズムを改変し，Web ページの推薦の問題に適用したものである．実験を通じて，このアルゴリズムが情報推薦問題で問題となるコールドスタート問題が発生するような条件下でよい推薦結果を示すことが分かった．閲覧履歴情報が極端に多い，少ない状況を仮定した場合の評価のみであるためより詳細な評価については今後の課題とする．

謝辞 本研究の一部は，科学研究費補助金基盤研究 (B)（23300039）および基盤研究 (A)（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．

参 考 文 献

1) 土方嘉徳：情報推薦・情報フィルタリングのためのユーザプロファイリング技術，人工知能学会誌，pp.365–372 (2004).
2) 土方嘉徳：嗜好抽出と情報推薦技術，情報処理，pp.957–1965 (2007).
3) HongboDeng, MichaelR.Lyu and IrwinKing: A generalized Co-HITS algorithm and its application to bipartite graphs, *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, ACM, pp.239–248 (2009).
4) M.Bishop, C.: パターン認識と機械学習 上ベイズ理論による統計的予測，シュプリンガー・ジャパン (2007).
5) Zhou, D., Bousquet, O., Lal, T.N., Weston, J. and Schölkopf, B.: Learning with local and global consistency, *Advances in Neural Information Processing Systems 16*, MIT Press, pp.321–328 (2004).
6) Zhou, D., Schölkopf, B. and Hofmann, T.: Semi-supervised learning on directed graphs, *In NIPS*, MIT Press, pp.1633–1640 (2005).
7) Zhu, X., Ghahramani, Z. and Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, *IN ICML*, pp.912–919 (2003).
8) Jamali, M. and Ester, M.: TrustWalker: a random walk model for combining trust-based and item-based recommendation, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, ACM, pp.397–406 (2009).
9) Massa, P. and Avesani, P.: Trust-aware recommender systems, *Proceedings of the 2007 ACM conference on Recommender systems*, RecSys '07, New York, NY, USA, ACM, pp.17–24 (2007).
10) D.Manning, C., Raghavan, P. and Schütze, H.: *Information Retrieval*, Cambridge University Press (2008).