社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ICS－147 (21)
2007／3／15

ユーザの網羅性を反映したランキング手法の提案

川前徳章　山田武士

NTT コミュニケーション科学基礎研究所〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: {kawamae.noriaki, yamada}@cslab.kecl.ntt.co.jp

あらまし 本研究ではユーザの検索活動を効率化するために、網羅性という観点からみたユーザの重要性を考慮したランキング手法を提案する。提案するランキング手法はユーザの網羅性の判定と Naive Bayes に基づいたランキングモデルから構成される。ここで重要なユーザは、検索対象となる特定分野のコンテンツに対して幅広く網羅的にアクセスしていると考える。更にユーザはランキングをユーザの網羅性の観点からパーソナライズすることが可能となる。提案手法を実験に用いた結果、ランキングに網羅性を反映できることとユーザの検索活動の効率化を確認できた。

キーワード ランキング 情報検索 パーソナライズ 協調フィルタリング コンテンツフィルタリング

Ranking Method Incorporating User Coverage

Noriaki KAWAMAE and Takeshi YAMADA

NTT Communication Science Laboratories 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
E-mail: {kawamae.noriaki, yamada}@cslab.kecl.ntt.co.jp

Abstract This paper proposes a novel ranking method incorporating user content coverage for improving user information retrieval activity. The proposed method is composed of the detection of users with high content coverage and the ranking model based on Naive Bayes. We assume that the import users can be identified by their comprehensive contents coverage of a target domain. The advantage of this method is to enable users to adjust and personalize ranking in view of coverage. We have applied the proposed method to a benchmark data set and confirmed that it enabled us to incorporate user content coverage in the ranking resulting in improving user information retrieval activity.

Keyword Ranking Algorithm, Information Retrieval, Personalization, Collaborative Filtering, Contents Filtering

1. はじめに

本研究の目的はユーザのコンテンツ検索を効率化することにあり、その実現に向けてコンテンツの新しいランキング手法を提案する。ランキング手法には *tf-idf* に基づいてコンテンツをランキングする手法[3]や Web ページのリンク構造を解析しランキングする HITS などの手法[5]がある。これらの従来手法には適用性とランキングの更新頻度という二つの問題が存在する。前者はランキングの対象となるのがコンテキストを含む情報かリンク構造があるものに限られるという問題である。後者はコンテキストもリンクもほぼ静的なものであり、実際のユーザの動的な興味や人気が反映されにくいという問題である。そこでこれらの問題点を解決するために、ランキング手法に協調フィルタリングを導入することを提案する。協調フィルタリングはユーザの履歴を利用するので、履歴に含まれるコンテンツであれば、コンテキストやリンク情報が無いコンテンツもランキングの対象になる。更に履歴には時間情報も含まれることから最新のユーザの行動を反映した動的なランキング手法が実現できると考えられる。

ランキングの予測精度の向上によってユーザの検索効率はより向上する。協調フィルタリングを用いたランキングの予測精度は利用するユーザの履歴の質に依存すると考えられる。協調フィルタリングはサーバのアクセスログやブラウザの閲覧履歴などのユーザの履歴をユーザ個人のみならずユーザ全体で共有し、ユーザの興味を推定し、そのユーザに合ったコンテンツを予測して提示する手法である。一方で、既存の協調フィルタリングの問題点は興味が同じユーザでも知識の量と網羅性が違う場合、この違いを利用してこなかったことである。例えば、技術資料を検索する場合を考える。その技術に関する初心者と網羅的な知識を持つ上級者とでは必要とする情報は異なるにも関わらず、興味が同じなら、協調フィルタリングの結果は同一となる。

本研究はユーザの網羅性を考慮した協調フィルタリングに基づくランキング手法を提案する。提案手法

はユーザ及び検索対象となるコンテンツ共に網羅性による違いがあるので、ランキングに網羅性が反映されることで、ユーザの検索活動が効率化されるというアイデアに基づく。このアイデアに従い、Naive Bayesを用いて、各ユーザの履歴の網羅性を判定し、その網羅性を協調フィルタリングに反映し、ユーザは自分の網羅性に合わせてカスタマイズできるランキング手法を提案する。Naive Bayesは多くのテキストフィルタリングのアプリケーションにも実装されているモデルで、情報検索でよく知られた *tf-idf* とも数学的に非常に関係が深い。提案するランキング手法は協調フィルタリングが確率モデルの観点から明快な対応関係があるという事実に基づいている。

従来の協調フィルタリングでは提示された結果に対してユーザがフィードバックをかけることはあまり行われていなかったのに対し、提案するランキング手法はユーザが自身の網羅性に合わせて結果を調整することが出来る。これはユーザが検索において自身の網羅性とは違った情報を検索するニーズがあると考えられ、そのニーズをランキングにフィードバックすることで提案手法のランキングモデルを活用できるというアイデアに基く。その結果、ユーザが自身の網羅性を考慮してランキングをパーソナライズすることで検索活動を効率化できると考える。実験において提案したランキング手法は他の従来手法に比較して予測精度を向上し、協調フィルタリングにおいても情報検索同様にNaive Bayesに基いたランキングモデルは精度向上に寄与することを確認した。

2. 既存研究

提案手法は協調フィルタリングを基にしたランキング手法である。これまでE-commerceから計算機上での作業支援まであらゆる文脈で多様な協調フィルタリングの手法が提案されてきた[6,8]。既存の協調フィルタリングはアルゴリズムの設計方針に応じてメモリベースかモデルベースのアプローチに分けられる[1]。アプローチの違いに拠らず、これらの手法は全ユーザの履歴集合を利用している。更に実際にはユーザには興味だけでなく知識の網羅性の違いも存在するにも関わらず、これらの手法は興味の違いのみをユーザ間の類似性として利用してきた。提案手法は従来手法のユーザの膨大な履歴集合の処理を特徴選択手法で軽減し、更にその特徴選択においてユーザの網羅性を考慮することで予測精度を向上することを目的とする。

3. ユーザの網羅性を考慮したランキング

3.1. ユーザの網羅性

ここで我々は検索対象の特定分野に関して他のユーザよりも良く知っている人を網羅性の高いユーザと呼び、その網羅性とそれを判定する方法を提案する。

我々はユーザの網羅性は、全ユーザの履歴集合に対する、そのユーザの履歴の情報損失度として評価できると考える。特徴選択の考え方にしたがって、網羅性の高いユーザを選択することによって、元々のデータの質を維持しつつ、サイズを小さくしノイズを除去し般化性能を上げることができる。ユーザ u の網羅性とは、ユーザの履歴 h_u によって全ユーザの履歴集合 H を近似できる確率として定量化できる。この確率が高い履歴 h_u はそれだけ履歴集合 H の持つデータの質を維持していると解釈できる。この定義に基づいて我々はNaive Bayesを用いてユーザ集合 U におけるユーザ u の網羅性を次のように定式化する。

$$w(u)=w(h_u)=p(h_u|H)\propto p(H|h_u)p(h_u)$$
$$=\prod_{d\in D}p(d|h_u)^{N_d}\,p(h_u) \quad (1)$$

ここで h_u はユーザ u の履歴、H は全ユーザの履歴集合、D は全ユーザの履歴集合 H に出現するコンテンツ集合、N_d は H における d の出現頻度、$P(h_u/H)$は h_u の H での出現確率、$P(d/h_u)$は d の h_u での出現確率、そして $P(h_u)$は履歴 h_u の出現確率である。$P(d/h_u)$は個々のユーザの履歴から、N_d 及び $P(h_u)$は全ユーザの履歴集合から推定される。$P(h_u)$ 及びゼロ頻度問題を回避する為に $P(d/h_u)$は *Jeffreys Perks Law* を用いて以下の式で推定する。

$$p(h_u)=\frac{1}{\text{全ユーザ数}} \qquad p(d|h_u)=\frac{0.5+n_{ud}}{0.5*|D_u|+N_u} \quad (2)$$

ここで n_{ud} は h_u における d の出現頻度、$|D_u|$は h_u におけるコンテンツの異なり数、N_u は h_u におけるコンテンツの出現頻度の合計である。式(1)が意味するのは、ユーザは履歴に含まれるコンテンツの出現確率の積によって特徴付けられると言うことである。履歴集合 H においてコンテンツ d の出現頻度 N_d が高ければそれだけ各ユーザの履歴 h_u における重みも高くなる。このことから多くのユーザによってアクセスされるコンテンツをより多くアクセスしているユーザは他のユーザに比較して $w(u)$ の値は高くなる。我々は式(1)による確率をユーザの網羅性として扱い、この網羅性の値が高い順に協調フィルタリングに入力するユーザの履歴を選択する。

式(1)は式変形により、Kullback Leibler divergenceに基づく情報損失に対応する[3]ことが分かる。式(1) の対数を取り、H のサイズで割ることで次のように変形できる。

$$w'(u) \equiv \frac{1}{|H|}\left(\log p(h_u) + \sum_{d \in H} N_d \log p(d|h_u)\right) \quad (3)$$
$$= \frac{1}{|H|}\log p(h_u) + \sum_{d \in H} p(d|H)\log p(d|h_u)$$

$P(d/H)$のエントロピーを引き、第一項の $P(h_u)$は全ユーザでほぼ等しいので無視し、第二項のみに着目することで式(3)は以下のように変形できる。

$$w'(u) = \frac{1}{|H|}\log p(h_u) + \sum_{d \in H} p(d|H)\log \frac{p(d|h_u)}{p(d|H)} \quad (4)$$
$$\equiv D(H|h_u) = D(U|u)$$

ここで $D(U/u)$はユーザ集合 U及びユーザ uの確率分布間の Kullback Leibler distance を示しており、式(4)の値を最大とすることで、ユーザ集合 Uに対して情報損失を最小にするユーザ uを選択することが出来る。式変形により式(4)が意味するのは式(1)において網羅性が高いと判断されるユーザ uの発見は全ユーザの履歴集合 Hに対して情報損失が最小となるユーザ uの履歴 h_uの発見と等価ということである。情報損失が最小の履歴は履歴集合を近似した履歴であると見ることができる。従って、我々はこの Naive Bayes に基づいたユーザ uの網羅性を検索対象の各分野におけるユーザの網羅性を評価するのに用いる。

3.2. 協調フィルタリング、情報検索及びテキストフィルタリングとの関係

ここでは協調フィルタリングとテキストフィルタリング及び情報検索の関係について説明し、ランキングモデルを提案する。テキストフィルタリングの目的は未分類のテキストに対してカテゴリのラベルを正しく付けることにある。情報検索の目的はユーザのクエリに関して関係あるテキストにランクを付けてユーザに対して最も適切なテキストを提示することにある。協調フィルタリングの目的はユーザの履歴を使うことで、コンテンツにランクを付けてユーザに対して最も適切なコンテンツを提示することにある。これら三つのタスクはカテゴリ(テキスト、コンテンツ)とテキスト(クエリ、履歴)の関係を測定するように設計され、以下のように定式化できる。

テキストフィルタリング:

$$\underset{c_k}{\operatorname{argmax}}\ p(c_k|d_j) \propto \underset{c_k}{\operatorname{argmax}}\ p(d_j|c_k)p(c_k) \quad \text{(a)}$$

c_k:カテゴリ　d_j:テキスト

情報検索

$$\underset{d_j}{\operatorname{argmax}}\ p(d_j|q) \propto \underset{d_j}{\operatorname{argmax}}\ p(q|d_j)p(d_j) \quad \text{(b)}$$

d_j:テキスト　q:クエリ

協調フィルタリング

$$\underset{d_j}{\operatorname{argmax}}\ p(d_j|h_u) \propto \underset{d_j}{\operatorname{argmax}}\ p(h_u|d_j)p(d_j) \quad \text{(c)}$$

d_j:コンテンツ　h_u:ユーザの履歴

図1. テキストフィルタリング、情報検索及び協調フィルタリングの確率モデル.

これらの式が意味するのは、テキストフィルタリング、情報検索そして協調フィルタリングのタスクは事後確率を最大化するカテゴリ(テキスト、コンテンツ)をユーザに結果として提示するということである。これらのモデルは本質的に事後確率最大化の問題になっている。それ故、協調フィルタリングがその目的だけでなく確率モデルとしてもテキストフィルタリング及び情報検索と明確な対応関係を持っていると言える。

ランキングアルゴリズムは Naive Bayes を用いた確率モデルとして考えることが出来る。Naive Bayes はテキスト生成の確率モデルであり、ベイズ則を用いてテキストのカテゴリを予測する問題と書き換えることが出来る。これはテキスト内の単語は独立で、テキストのカテゴリはベイズ則を用いて単語のカテゴリにおける出現確率とテキストにおける出現頻度の積によって予測されるというアイデアに基づいている。ここで我々はこの Naive Bayes を協調フィルタリングに適用し、ユーザ uに対するコンテンツのランキングモデルを次のように定義する。

$$\underset{d_i}{\operatorname{argmax}} P(d_i|u) = \underset{d_i}{\operatorname{argmax}} P(d_i|h_u) \propto \underset{d_i}{\operatorname{argmax}} P(h_u|d_i)P(d_i)$$
$$= \underset{d_i}{\operatorname{argmin}}\left(\sum_{d_j \in h_u} n(d_j)\log P(d_j|d_i) + \log P(d_i)\right) \quad (5)$$

先の定義より網羅性の高いユーザはそうでないユーザが知らないコンテンツを知っていると考えられる。そこでこのランキングモデルにそのユーザの網羅性を反映させる為に、情報検索における異なる確率分布を用いたスムージング[10]に従い $P(d_j|d_i)$を $\lambda * P_s(d_j/d_i)+(\lambda-1)*P(d_j)$とスムージングすることで式(4)を以下のように変形する。

$$\operatorname*{argmax}_{d_i} P(d_i|u)$$

$$= \operatorname*{argmin}_{d_i} \left(\left(\sum_{d_j \in h_L} n(d_j) \left(\lambda * \log P_s(d_j|d_i) + (1-\lambda) * \log P(d_j) \right) \right) + \log P(d_i) \right)$$

$$where \quad P_s(d_j|d_i) = \frac{|U_{sij}|}{|U_{si}|} \qquad (6)$$

ここで $n(d_j)$は d_j が h_u に出現する確率、$P(d_i)(P(d_j))$はユーザ集合がコンテンツ集合から $d_i(d_j)$を選択する確率、$P_s(d_j|d_i)$は d_i を選択した網羅性の高いユーザ集合が d_j を選択する事後確率、$|U_{si}|$は d_i を選択した網羅性の高いユーザの数、$|U_{sij}|$は d_i 及び d_j を選択した網羅性の高いユーザの数、そして λ は[0,1]の範囲のパラメタであり、全ユーザにおける網羅性の高いユーザの重みを決定する。

ユーザの網羅性の判定とそれを用いたユーザ選択の目的はランキングモデルのパラメタの推定とそのスムージングにある。従来手法がパラメタ推定に全ユーザの履歴集合を用いるのに対して、我々はユーザを式(1)の値によってユーザの網羅性を判定し、その結果、網羅性の高いユーザの履歴のみを用いて $P_s(d_j|d_i)$を推定する。

パラメタ推定とスムージングの効果はデータスパースネスや過学習の問題の回避にある。式(5)に示したように、ランキングモデルはユーザの履歴において同時に出現したコンテンツの条件付確率の積になっている。しかしながら、個々のユーザの履歴で見た場合、これらの確率は 0 になっており、データスパースネスの問題がある。ユーザの網羅性の定義から、ユーザの網羅性を考慮して推定した $P_s(d_j|d_i)$はユーザの網羅性を考慮せずに推定した $P(d_j)$よりも高くなる傾向がある。従って、ユーザの網羅性を考慮して推定した条件付確率 $P_s(d_j|d_i)$はそうでないユーザに対して知らないコンテンツを気が付かせるだけでなく、一方で $P(d_j)$を用いてスムージングすることで過学習を避けることができる。

更にこのランキングモデルの特徴はユーザがパラメタλを調整することでユーザがランキングに対してフィードバックすることが可能となる点にある。その結果、ユーザはパラメタの調整によりランキングを自身の網羅性に合わせ、あるいはコンテンツの網羅性を高低することでパーソナライズすることが可能となる。

4. 実験

4.1. 実験に用いたデータ

我々は今回の実験で Each Movie データセットを用いた。Each Movie data set は Digital Equipment Corporation(現在 Hewllet-Packard)が運営していた映画の協調フィルタリングのサイトから作ったデータである。このデータは 72916 人の利用者の 1628 種類の映画に対する評点から構成されていて、評点は 0-1 の範囲で 0.2 の間隔尺度でされている。このデータセットは一般に利用できるデータセットの中で最も規模が大きく、利用者の評点を含んでいるという特徴を持つことから、協調フィルタリングの手法を評価するベンチマークデータとして利用される[7]。

今回の実験の目的は各種ランキングモデルによる top-N リコメンデーションの結果を質とパフォーマンスの点で評価することである。top-N リコメンデーションの結果を質の面から評価する為に、我々はデータセットからユーザ毎に評点が 0 でない映画を任意に抽出してデータの 25%をテストデータ、残り 75%を学習データとして分けることを全ユーザについて行うことを 10 回繰り返した。各ユーザに対してランダムに 1 から 10 までの番号を振り、n 回目の繰り返しの時は、n の番号を付けられたユーザのデータは学習データから除く。残った学習データはパラメタの推定に用い、テストデータは学習データに含まれないユーザのランキングの観測値として用いた。本実験の結果はこの 10 組のデータに対して行った平均である。

4.2. ユーザの網羅性を考慮しないランキングモデル

最初の実験において、提案するランキングモデルにおけるスムージングの効果を modified precision[7]を用いて評価する。特徴選択の効果を比較する為に今回の実験においては式(5)の $P_s(d_j|d_i)$を網羅性の高いユーザでなく全ユーザの履歴集合を用いて推定した。図 2 において *top-N* リコメンデーションのリスト N の大きさとパラメタλの値を変化させた時の modified precision を示す。図 2 の結果から top-1 及び 5 の modified precision はλの影響を受け無いことが分かる。有意水準 95%の両側 t 検定を用いて統計的な有意差は見られなかった。しかし、top-10 及び 20 においてはλの値によって変化が見られた。この変化はλが 0.2 から 0 へと変化する時に減少し、この差は有意水準 95%の両側 t 検定で有意だった。この実験結果から top-N リコメンデーションの N が大きくなると、スムージングの効果が modified precision に出てくる。

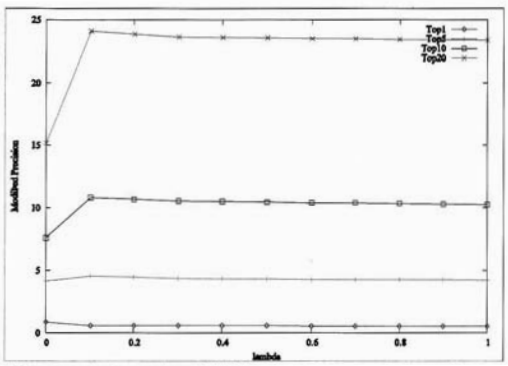

図 2.　スムージング効果による top-N のサイズ毎の modified precision の変化

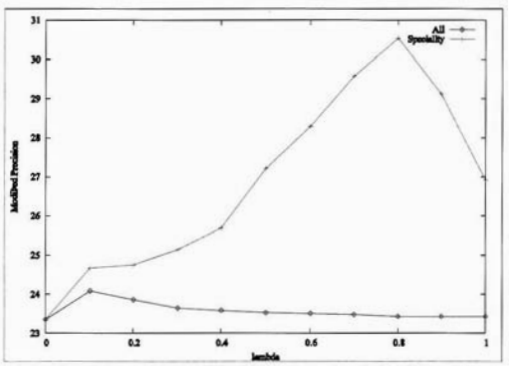

図 3.　スムージング効果とユーザの網羅性の有無による modified precision の変化

4.3. ユーザの網羅性を考慮したランキングモデル

この実験では、ランキングにおけるユーザの網羅性の効果を前回同様に modified precision を用いて評価する。今回の実験ではベンチマークとして前回の実験の top-20 リコメンデーションを用い、式(5) における $P_s(d_j|d_i)$を網羅性の高いユーザの履歴集合から推定し、$P(d_i)(P(d_j))$を全ユーザの履歴集合から推定することで得られる top-20 リコメンデーションの結果と比較した。パラメタ推定において網羅性の高いユーザは式(3)の値によって上位 5 %のユーザを選択した。図 3 において top-N リコメンデーションのリスト N の大きさとパラメタ □ の値を変化させた時の modified precision を示す。図 3 の結果から前回の実験と異なり modified precision は □ の影響を受けることが分かる。□ が 0 から 0.8 までは modified precision が増加し、それ以降は減少している。3.2 で示したように □ はモデルにおける $P_s(d_j|d_i)$の割合、つまり網羅性の高いユーザの履歴がランキングに占める割合である。この実験において □ が 0.5 から 0.6 及び 0.9 から 1 の間では有意水準 95% の両側 t 検定、□ が 0.6 から 0.9 の間では有意水準 99% の両側 t 検定で有意差がある。□ の最適な値が 0.8 周辺であることは履歴のデータスパースネスの影響であると考えられる。データスパースネスの為に □ が 0.8 まで増加するまでは $P_s(d_j|d_i)$の値は $P(d_j)$よりも低く、網羅性の高いユーザの履歴がランキングに反映されなかった為と考えられる。

更に我々は網羅性の高いユーザの嗜好を評価する為にテストデータに出現する映画を層別して modified precision の変化を調べた。まず、履歴集合においてアクセスユーザ数が多い順に Top100、少ない順に Bottom100 の映画を選択し、次に各 λ に対して、テストデータに含まれるそれぞれの映画の modified precision を調べた。図 4 にその結果を示す。

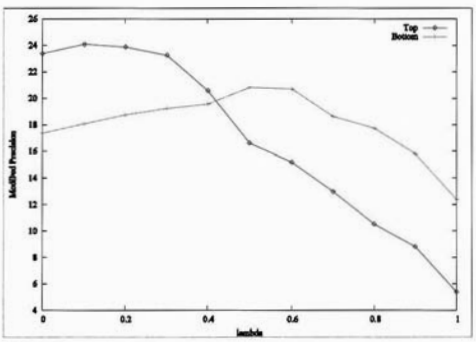

図 4.　スムージング効果と映画の人気の有無による modified precision の変化

図 4 において Top、Bottom 共に λ が 0.6 以上では modified precision は減少傾向を示している。この理由は網羅性の高いユーザは他のユーザには評価されないような映画も評価しているものの、そのような映画がテストデータに含まれる数が少ない為と考えられる。表 1 に Top10 及び Bottom10 の映画のタイトルを示す。この表における()内の数値はデータ集合の中でアクセスしたユーザの数である。

表 1．Top、Bottom の映画のタイトル

Top 10	Bottom 10
Batman (1989) (32864)	Sunday (1)
Dances With Wolves (32674)	Suicide Kings (1)
Apollo 13 (31259)	Fire (2)
Pulp Fiction(30510)	Gravesend (2)
True Lies (29378)	A Self-Made Hero (3)
Ace Ventura: Pet Detective (25990)	The Toilers and the Wayfarers (3)
Aladdin (29376)	Trojan Eddie (3)
The Fugitive (24279)	Julian Po (3)
Batman Forever (24260)	Moonlight Murder (4)
Die Hard: With a Vengeance (24000)	The Winner (4)

この表からもランキングモデルにおいて λ の上昇と共に modified precision が減少した理由を考えることが出来る。Bottom の映画はユーザに評価されにくく、結果

として履歴の中での出現頻度も少ない。従って $P(d_i)$ の値が他の人気のある映画よりも低くなり、式(5)のランキングモデルにおいて上位にランクされにくくなる。一方でλの値を高くすることによって、提案するランキングモデルは他のユーザにはなかなか評価されない映画のランクを上げることが出来、ユーザにとって自分では気が付かない新たな映画を発見することに繋がると期待できる。

4.4. 従来手法との比較

ここではmodified precisionを比較する為に、典型的なランキングモデルのベンチマーク手法としてrelevance model [9], the Belief Distribution Algorithm [7] , Okapi [4], Naive Bayes model においてパラメタをLaplace,Jeffreys-Perks Lawsを用いて推定を行ったモデル、mutual information, conditional entropyを用いた。これらの手法は情報検索や協調フィルタリングにおいてその効果を知られているものである。これらの手法のmodified precisionをtop-NリコメンデーションのリストNの大きさを変化させた場合の結果を図5に示す。

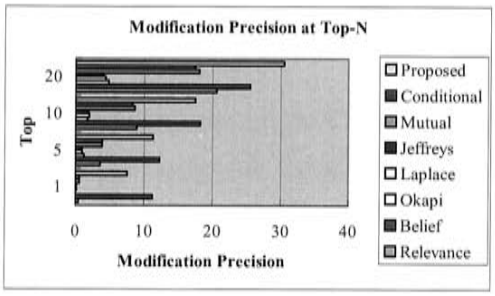

図 5. Relevance model (Relevance), Belief Distribution Algorithm (Belief), Naive Bayes with Laplace's Law (Laplace), Naive Bayes with Jeffreys-Perks Law (Jeffreys), mutual information (Mutual), conditional entropy (Conditional), Okapi (Okapi) そして提案手法 (Proposed) による top-N のサイズ毎の modified precision の変化

図5の結果から、我々はBelief と提案手法が他の手法に比べてmodified precisionが高いことが分かる。一方でこの二つの手法の優劣はNの大きさによって変化している。Belief はtop-1 において有意水準95% で、提案手法はtop 20 において有意水準99% で、両側t検定を用いて有意であることを示している。top-5及び20 では統計的に有意な差は見られなかった。

5. 考察

実験結果は協調フィルタリングにおけるパラメタ推定の網羅性についても示している。Okapi は情報検索において効果があると知られている手法であるが、今回の実験ではその効果が見られなかった。これは履歴におけるスパースネスの問題が他の分野に比較して顕著であり、その結果、パラメタ推定の違いが予測精度に影響を与えることを示していると考えられる。以上の実験を通してユーザの網羅性の着目したユーザ選択手法はパラメタ推定にとって必要な処理であり、Naive Bayes に基づいたランキングモデルは情報検索同様に協調フィルタリングでもその効果を発揮することを確認した。また興味だけでなく網羅性の違いも協調フィルタリングの精度向上に重要であることを確認した。

6. 結論

提案手法の協調フィルタリングにおける貢献はユーザの網羅性を考慮したランキングによって予測精度を向上したことにある。提案したランキングアルゴリズムはランキングの精度向上の為に入力においてユーザの特徴選択、処理において Naive Bayes モデルの利用、出力においてユーザフィードバックを可能とした。実験において提案手法は予測精度の向上を実現し、その有効性を確認することが出来た。

文　献

[1] J. S. Breese, D. Heckerman and C. M. Kadie: Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, pp. 43-52, July 1998.

[2] M. Deshpande and G. Karypis: Item-based top-N recommendation algorithms. ACM Trans. Inf. Syst. 22(1), pp. 143-177, 2004.

[3] I. S. Dhillon, S. Mallela, R. Kumar: Enhanced word clustering for hierarchical text classification. KDD 2002, pp. 191-200, 2000.

[4] K. S Jones, S. Walker and S. Robertson: A probabilistic model of information retrieval: development and status. University of Cambridge Computer Laboratory Technical Report no. 446, 1998.

[5] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5):604?632, 1999.

[6] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl. GroupLens: Applying collaborative filtering to Usenet news. Communications of the ACM, 40(3), pp. 77-87, 1997.

[7] M. R. McLaughlin and J. L. Herlocker: A collaborative filtering algorithm and evaluation metric that accurately model the user experience. SIGIR 2004, pp. 329-336, 2004.

[8] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommender algorithms for e-commerce. In Proceedings of the 2nd ACM Conference on Electronic Commerce, pp 158-167, 2000.

[9] J Wang, A. P. d. V. and M. J. T. Reinders: A user-item relevance model for log-based collaborative filtering, ECIR-2006, 2006.

[10] C. X. Zhai and J. D. Lafferty: A study of smoothing methods for language models applied to ad hoc information retrieval. SIGIR 2001, pp. 334-342, 2001.