

投稿型サイトにおけるユーザの興味と流行 を考慮したチャート型ランキングの個人化 手法

金子鷹弥[†] 牛尼剛聡^{††}

現在, Web 上には多くの投稿型サイトが存在し, 一般のユーザもコンテンツを閲覧するだけでなく, 簡単にコンテンツを投稿することが可能になった. それに伴い, 投稿型サイトに投稿されるコンテンツ数は莫大になっている. ユーザは日々大量に投稿される大量のコンテンツの中から, アクセスに値するコンテンツを選別する必要がある. しかし, 新規に投稿されたコンテンツは, ユーザにとって未知のコンテンツであり, 効率的な選別が困難である. そこで, 我々はチャート型ランキングに着目した. 一般的に, チャート型ランキングは流行度を基準としてコンテンツを順位付けているが, 提案手法では, 流行度の他にユーザの嗜好を取り入れることにより, 投稿型サイトにおけるユーザの適切なコンテンツアクセスを支援する.

A Method for Personalizing of Chart Ranking Based on Popularity and User Preferences for Shared Content

TAKAYA KANEKO[†] TAKETOSHI USHIAMA^{††}

Today, many Web services that are used to share contents exist and users become not only a content consumer but a contributor. Therefore, the number of contents are posted to content sharing sites is enormous. The user would look for the contents that interest the user to access from the contents contributed every day. However, the features of new contents have not been known by the user, so the it is difficult for the user to select suitable contents from them with the existing technique. In this paper, we introduce a method to support user's content access for shared content by personalizing chart ranking based on their popularity and user preferences on them.

1. はじめに

近年, インターネットの普及と Web 上のコンテンツの増加に伴い, ユーザのコンテンツの利用形態が変化している. インターネットが普及する以前は, テレビ放送やラジオ放送のようなブロードキャスト型のコンテンツ配信が広く利用されていた. ブロードキャスト型のコンテンツ配信では, 多数のユーザに対して同一のコンテンツを同時に配信する. ブロードキャスト型のコンテンツ配信では, ユーザはコンテンツを選択する必要がない代わりに, 興味の無いコンテンツを閲覧する可能性が高く, ユーザの満足度は高くない. これに対して, 近年急速に普及しているインターネットを利用したコンテンツ閲覧では, ユーザが興味のあるコンテンツを発見し, 好きな時に閲覧することができるため, ユーザの満足度を向上させることが可能である.

従来から利用されてきた Web 上での代表的なコンテンツアクセス手法として検索がある. 検索では, ユーザが検索対象の明確な特徴を想定し, クエリとして表現することができれば, ユーザが満足する可能性が高いコンテンツを高精度で提示できる. しかし, いくつかの状況に於いて, 検索では効果的なアクセスを提供できない場合がある. 例えば, 未知のコンテンツ集合に対しては, 検索によってユーザが見る価値のある発見することは難しいことがある. これは, 未知コンテンツに対しては, 検索対象の特徴が明らかになっていないため, クエリの生成が困難であるからである. そこで, 我々は, 未知のコンテンツを有効にユーザに提示する手法の開発を目標として研究を行っている.

未知のコンテンツが多数存在するものとして対象の代表例として, 投稿型サイトが挙げられる. 現在, 様々な Web サービスの登場により, Web 上に存在するコンテンツ数が膨大になっている. その大きな要因となったのが, YouTube やニコニコ動画はじめとする各種投稿型サイトの出現である. 投稿型サイトにおいて, ユーザは, コンテンツの閲覧者であると同時に, 発信者でもある. そのため, 投稿型サイトには, 短期間に膨大なコンテンツが投稿される. 投稿型サイトのユーザは, そのような膨大なコンテンツから, 閲覧する価値のあるコンテンツを選別する必要がある. また, 近年多くのユーザを獲得しているソーシャル・ネットワーキング・サービス(SNS)や Twitter をはじめとしたマイクロブログも, テキスト, 画像, 動画等, コンテンツの種類は違うものの, ユーザが投稿し膨大なコンテンツが生成されるという意味で, 投稿型サイ

[†]九州大学芸術工学府
Graduate School of Design, Kyushu University
^{††}九州大学芸術工学研究院
Faculty of Design, Kyushu University

トの一種と捉えることができる。

投稿型サイトにおいては、個々のユーザが、投稿された全てのコンテンツを閲覧することは困難である。そのため、投稿型サイトにおいて、ユーザは興味を持つコンテンツを様々な方法で探している。しかし、投稿型サイトに日々投稿されるコンテンツ数は膨大であり、その中からユーザの興味に適したコンテンツを的確に発見することは困難である。

投稿型サイトでは、投稿された多くのコンテンツをユーザに有効的に提示するために、アクセスに対するブックマーク数等のような特殊な評価値を用意する場合や、情報推薦等の様々なアクセス手法を提供している。

1.1 投稿型サイトにおけるコンテンツ発見

投稿型サイトを含む多くのコンテンツ提供サービスにおいて、最も頻繁に利用されるコンテンツのアクセス手法は検索である。検索は、ユーザが検索結果が有する特徴をクエリとして明示的に指定し、希望するコンテンツを取得する手法である。検索は、ユーザが検索結果となるコンテンツの特徴を想定できる場合には、効果的である。一般的に、ブロードバンド放送のように、ユーザが受動的にコンテンツを閲覧しなければならない場合には、興味のないコンテンツをも閲覧しなければならない。しかし、検索によるコンテンツ・アクセスでは、ユーザは主体的にコンテンツを選別するため、興味のないコンテンツを閲覧する必要がない。これは、ユーザの満足度の向上につながる。

しかし、検索では効果的なアクセスを提供できない場合も存在する。例えば、検索では、未知のコンテンツに対しては、ユーザはそれらの特徴を明示的にクエリとして指定することができない。そのため、ユーザは自分が興味のないコンテンツを閲覧しなくてよいという利点がある一方で、ユーザの興味を引く、ユーザが未知のコンテンツを発見することが難しくなっている。そのため、投稿型サイトにおいては、日々の新規投稿コンテンツが膨大であることから、それらを検索により発見することは困難である。

検索の欠点を補うアクセス手法として、近年、様々な情報推薦手法が研究されている。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できる。Grouplens[1]に代表される協調フィルタリングは、ユーザが興味を持つ可能性の高いコンテンツの提示が可能という点で、有効な手法であり、広く利用されている。強調フィルタリングは、対象とするユーザと嗜好が似ている別のユーザの情報を利用してコンテンツ提示を行う。しかし、利用初期や新規に投稿されたコンテンツを対象とする場合に、データが少ないことによるコールドスタート問題が発生する。毎日の投稿数が多い投稿型サイトにおいて、新規コンテンツに対しての正確な評価が難しいため、協調フィルタリングは新規コンテンツを対象とした推薦には不適切である。一方、内容ベースフィルタリングでは、ユーザのプロファイルに合致したコンテンツを推薦

る。しかし、ユーザのプロファイルが正確に取得できたとしても、ユーザが興味を持つ全てのコンテンツがプロファイルに合致しているとは限らないという問題がある。例えば、2011年7月に女子サッカーに興味を持ったユーザは、女子サッカーに対する興味ではなく、流行していたことが理由でアクセスした場合が考えられる。つまり、現在のユーザのプロファイルに合致していると言うだけではなく、そのときに流行しているコンテンツを提示する手法が有効な場合がある。

本論文では、チャート型ランキングをユーザのアクセス履歴に基づいて個人化することにより、新規コンテンツが多数投稿される投稿型サイトに於けるコンテンツの選別に有効な、コンテンツ提示手法を提案する。

2. 投稿型サイトに適したコンテンツ提示手法

毎日大量のコンテンツが投稿される投稿型サイトに適した、コンテンツ提示手法の開発を開発するために、「ユーザの興味」と「流行」という2つの視点に注目した。一般的に、ユーザは、主に以下の2つの観点から、Web上でのコンテンツの選別を行っていると考えられる。

(1) 自分の興味があるコンテンツの取得

(2) 一般的に関心をもたれている（流行している）コンテンツの取得

上記の2つの場合では、ユーザの行動は大きく異なる。(1)の場合、ユーザは検索や情報推薦等を用い、自分にとって有益なコンテンツへの絞り込みを行う。この場合、ユーザの嗜好に合わせた絞り込みを行うことで、ユーザは興味のないコンテンツを閲覧する必要がない。しかし、適切な絞り込みの為には、ユーザが目的とするコンテンツに対する明確なイメージを持っている必要がある。それに対して、(2)の場合、ユーザはブロードキャスト放送を見るかのように、興味のあるコンテンツ、興味のないコンテンツの区別をせず、閲覧している。この場合、興味による絞り込みがないため、ユーザによっては興味のないコンテンツばかりを閲覧することになる場合もある。

そこで、我々はユーザの興味と流行をともに考慮したコンテンツ提示手法を開発することを目標とした。コンテンツの価値とは、ユーザにとって、自身の現在の興味を反映しているかだけでなく、それを見ることによって自分の興味が喚起される可能性の高さ、または一般的な重要性など様々な視点から考えることができる。投稿型サイトにおいて、コンテンツの価値を評価するために、チャート型ランキングに注目した。チャート型ランキングでは、多くの利用者が視聴したコンテンツは価値の高いコンテンツであるという仮定の下、コンテンツのアクセスを行う。また、投稿型サイトの利用形態として、定期的・継続的にアクセスし、興味深いコンテンツを見ることができないか、ニュースを確認するようにアクセスする利用形態が考えられる。このような利用形態では、デイリー（1日単位）、ウィークリー（1週間単位）といった間隔で集計

され、網羅性のあるチャート型ランキングが有効である。

2.1 チャート型ランキング

アクセス数に基づいてコンテンツを順位付けたものを「チャート型ランキング」と呼ぶ。チャート型ランキングは従来から、一般的な流行に基づいて、コンテンツを推薦する手法として広く利用されている。チャート型ランキングの例を図1に示す。チャート型ランキングは、現在、多くの投稿型サイトにおいて利用されており、特にYouTubeやニコニコ動画では、アクセス数の他にも様々な評価基準によるランキングが提供され、ユーザの満足度を向上させている。

チャート型ランキングの最も大きな特徴として、ユーザの入力が必要ないことが挙げられる。そのため、コールドスタート問題が無く、新規コンテンツの提示が可能のため投稿型サイトでのコンテンツ提示に適している。また、チャート型ランキングはデイリー、ウィークリー等、複数の時間間隔で集計され、ユーザはサイトを利用する時間粒度の違いに基づいて、異なる時間間隔でのランキングを利用する。そのため、常に新しいコンテンツが提示可能であり、またユーザが未知のコンテンツに対しても明確な評価値を決定し、効果的なコンテンツ推薦を実現可能である。しかし、ユーザの嗜好を考慮していないため、現在のチャート型ランキングでは、ユーザが興味を持たないコンテンツが多々提示される。そこで、チャート型ランキングをコンテンツ推薦の手法とし、チャート型ランキングに対して、ユーザの興味を利用した個人化を行う手法を提案する。



図1：投稿型サイトに於けるチャート型ランキングの例 2), 3)

3. チャート型ランキングの個人化手法

我々が提案するチャート型ランキングの個人化手法の概要を図2に示す。本手法では、投稿型サイトの全ユーザのアクセスにより形成されたチャート型ランキングもとに、それをユーザに適したように個人化する。ユーザは提示されたチャート型ランキングに対し、自分が興味を持ったコンテンツにアクセスする。その際、ユーザがアクセスしたコンテンツ集合に対し、ユーザが興味を持ったコンテンツであると判断する。そして、取得されたチャート型ランキング閲覧の履歴により抽出したユーザの嗜好をフィードバックし、次回に提示するチャート型ランキングに反映させる。この処理を継続することにより、段階的にユーザの嗜好を反映したチャート型ランキングを出力できるようにする。また、ユーザがコンテンツにアクセスする際に、ユーザが過去のアクセスに於いて、流行をどの程度重要視しているかを判断し、個人の嗜好よりも一般的な流行を重要視するユーザにコンテンツ推薦を行う際には、ユーザの嗜好とはそれほど合致していない場合でも、流行度が高いコンテンツは、上位にランキングする。

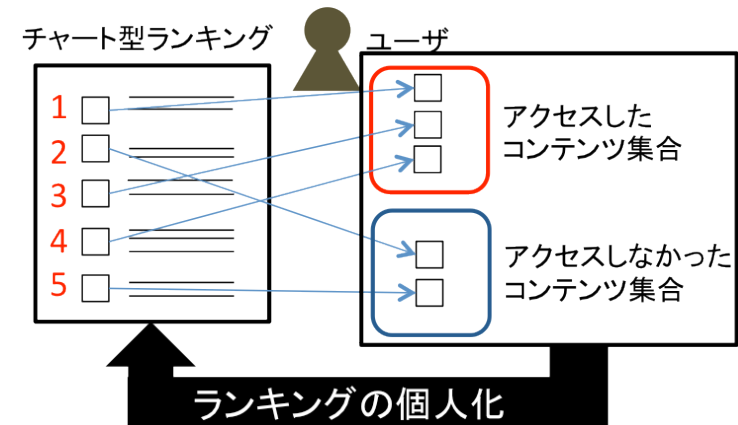


図2：提案手法の概要

3.1 ランキング手法

我々は、チャート型ランキングを個人化するために、確率モデルを利用する。具体的には、ランキングの対象となるコンテンツ集合に含まれる、あるコンテンツ c がユ

ーザに提示された場合に、ユーザがそのコンテンツをアクセス（閲覧）する条件付き確率 $P(\text{アクセス}|c)$ を求める。この確率を求める式は、ベイズの定理により以下のように変形できる。

$$P(\text{アクセス}|c) = \frac{P(c|\text{アクセス})P(\text{アクセス})}{P(c)}$$

いま、対象とするコンテンツ集合を C とすると、 C の構成要素全てに対して上記の条件付き確率を計算し、値が高いコンテンツから順にユーザに提示する。この式を計算する際、 $P(\text{アクセス})$ は、全てのコンテンツに於いて共通であるため、ランキングに影響は与えない。

ここで、我々は時刻 t におけるコンテンツ c_t を、コンテンツ c_t の特徴を表すキーワード集合 $\mathbf{w} = \{w_1, \dots, w_n\}$ と、アクセス数に基づいたランキング順位 r の組 $c_t = (,)$ としてモデル化する。 c_t の例を以下に示す。

$$c_1^t = (\{\text{なでしこ, サッカー}\}, 5)$$

$$c_2^t = (\{\text{AKB48}\}, 7)$$

ここで、 c_1^t は、キーワードが「なでしこ」、「サッカー」で、順位が 5 であるコンテンツを表す。一方、 c_2^t は、キーワードが「AKB48」で、順位が 7 であるコンテンツを表している。

上記のコンテンツ c_t のモデルを用い、 $P(c^t|\text{アクセス})$ は以下の式で計算できる。

$$P(c^t|\text{アクセス}) = P(\mathbf{w}|\text{アクセス}) \times P(r|\text{アクセス})$$

また、 $P(c)$ は以下のように計算できる。

$$P(c) = P(\mathbf{w}) \times P(r)$$

ここで、

$$P(\mathbf{w}) = \prod_{w_i \in \mathbf{w}_t} P(w_i) \times \prod_{w_i \notin \mathbf{w}_t} (1 - P(w_i))$$

である。

このとき、アクセスしたコンテンツの特徴語集合が $\mathbf{W}$ である確率 $P(\mathbf{w}_t | \text{アクセス})$ は以下の式で計算できる。

$$P(\mathbf{w}_t | \text{アクセス}) = \prod_{w_i \in \mathbf{w}_t} P(w_i | \text{アクセス}) \times \prod_{w_i \notin \mathbf{w}_t} (1 - P(w_i | \text{アクセス}))$$

ここで、ユーザ u がアクセスした全てのコンテンツ集合を $AC(u)$ 、単語 w を含むコンテンツ集合を $C(w)$ とすると、 $P(w_i | \text{アクセス})$ は以下の式によって求めることができる。

$$P(w_i | \text{アクセス}) = \frac{|AC(u) \cap C(w_i)|}{|AC(u)|}$$

図 3 に上記の計算を行うために必要な集合の関係を示す。本研究では、例題として動画コンテンツを対象とし、単語はコンテンツに付与されたタグを利用する。

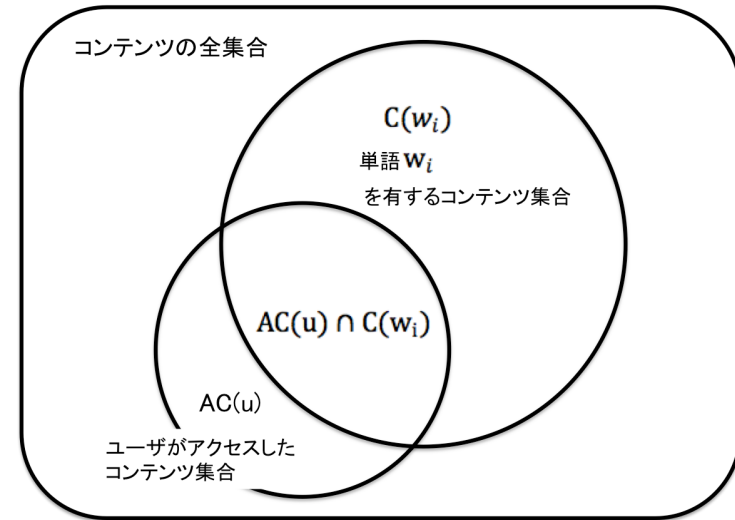


図3: $P(w_i | \text{アクセス})$ を計算するための集合の関係

また、ランクが r であるコンテンツの集合を $RC(r)$ とすると、アクセスしたコンテンツのランキングが r である確率 $P(r | \text{アクセス})$ は、指数分布に基づく以下の確率分布に従うものとする。

$$\begin{aligned} P(r | \text{アクセス}) &= \int_{r-1}^r \lambda e^{-\lambda x} dx \\ &= -e^{-\lambda r} + e^{-\lambda(r-1)} \\ &= (e^{-\lambda} - 1)e^{-\lambda r} \end{aligned}$$

このとき λ は標本平均の逆数で求められ、

$$\lambda = \frac{n}{\sum x_i}$$

となる。ここで、 x_i はユーザ u が i 番目にアクセスしたコンテンツのランキングである。 λ の値によって、 $P(r | \text{アクセス})$ の値は変化する。 λ の値が1, 0.2, 0.1であるときの、それぞれの $P(r | \text{アクセス})$ の分布を図4に示す。

我々の手法では、 $P(w_t | \text{アクセス})$ によって、ユーザの嗜好に合ったコンテンツを評価する基準を与え、 $P(r | \text{アクセス})$ により、上位のコンテンツに多くアクセスするユーザに対しては、個人化を行う以前のチャート型ランキングを重視し、流行しているコンテンツを提示するための流行を重要視する基準を示す。ユーザによっては、とにかく自身の嗜好にあったコンテンツに対して多数アクセスしたい場合や、流行しているコンテンツのみにアクセスしたい場合等、様々な要求が存在する。本研究において個人化を行う際、前述の $P(c^t | \text{アクセス})$ 、 $P(r | \text{アクセス})$ の値により、ユーザに適した個人化を行うことができる。

本手法では、ユーザがコンテンツにアクセスする度に $AC(u)$ や $RC(i)$ が変更されるため、ユーザがシステムを使って行くにつれて、ユーザの興味のある事項と、流行への順応の度合いを考慮してランキングが行われる。

我々が想定するチャート型ランキングの例を、図5に示す。本手法では、流行を意識しないユーザに対しては、 $P(r | \text{アクセス})$ の値が小さくなるため、個人化以前のチャート型ランキングにおける順位があまり影響せず、ユーザの嗜好を表す $P(c^t | \text{アクセス})$ の値により、主にコンテンツの評価が決定される。つまり、個人化後にはユーザの嗜好として抽出した特徴語を含むコンテンツが上位にランキングされる。また、流行への意識が強いユーザに対しては、 $P(r | \text{アクセス})$ の値が上位であるほど

高くなるため、上位のコンテンツは個人化以前のように高く評価される。ただし、ユーザの嗜好と合致するコンテンツはより高い評価を受ける。

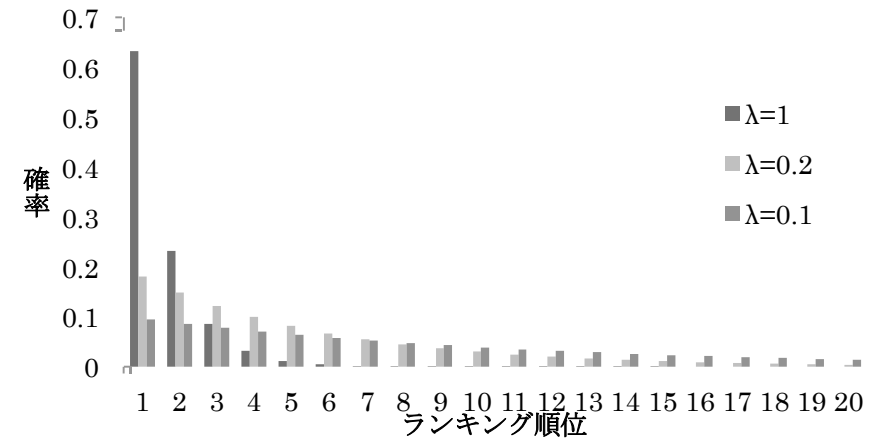


図4：様々な λ の値に対する $P(r | \text{アクセス})$ の分布

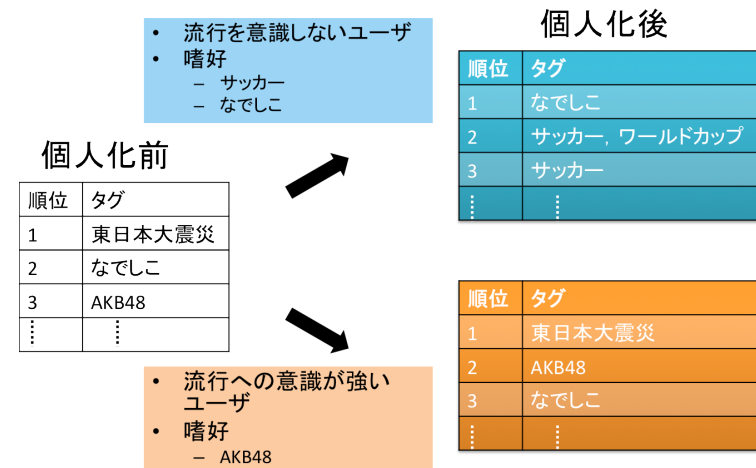


図5：本手法を用いた個人化の例

4. プロトタイプシステム

提案手法の有効性を検証するために、プロトタイプシステムを実装した。プロトタイプシステムの実行例を図6に示す。この画面では、ランキングされたそれぞれのコンテンツに関して、順位、タイトル、動画サムネイル、タグが示されている。本システムでは、ユーザの利用開始時には、一般的な全ユーザのアクセス数に基づく流行に基づくチャート型ランキングを提示する。ここでは、YouTubeで公開されているチャート型ランキングを利用している。YouTubeのチャート型ランキングは2時間毎に更新されており、我々のシステムにおいても、同様の間隔でチャートは更新可能である。現在、提示されるコンテンツは上位20件であり、ユーザはチャート型ランキング中のコンテンツのサムネイルをクリックすることにより、クリックした動画コンテンツを視聴ができる。

プロトタイプにおいて、数週間ユーザの動画の視聴ログを収集した。対象となるユーザは2名である。ここで収集したログに対して、それぞれのチャート型ランキングの順位における、アクセス頻度を図7に示す。また、ここで得られたデータを標本データとして、 $P(r | \text{アクセス})$ のパラメータを推定した。このとき、 λ の値は0.425531915だった。

この値を用いた場合の、 $P(r | \text{アクセス})$ の分布を図8に示す。提案手法によって得られたランキングの有効性を、被験者実験によって評価する予定である。

5. 関連研究

ランキングの研究は主に検索エンジンが提示する結果を再ランキングするというものが多く、提案手法のようにアクセス数ベースのチャート型ランキングを用いるものは少ない。山本ら[5]は、ランキング結果をユーザの様々なインタラクションにより、自在に変化させることが可能なRerankEverythingを提案した。RerankEverythingは本研究と同様に、チャート型ランキングをユーザの意思で変化させることが出来るが、操作を加えるのはユーザであり、我々が対象とした投稿型サイトにおける未知の新規コンテンツを発見することは難しい。

小林ら[6]は、仮想の生態系を用いることにより、情報検索を行うInformation Fishingを提案している。Information Fishingはクエリを餌、被検索文書を魚とし、ユーザの操作によって自在に変化する検索結果の提示を行うことができる。また、提示されるコンテンツには、ユーザの操作次第で未知のコンテンツが含まれる場合もあり、投稿型サイトにおいて適用することも十分可能である。しかし、クエリを入力する必要がある、未知のコンテンツの発見に対して、十分な機能を提供しているとは言えない。

湯本ら[7]が提案するWebページ集合を解とするランキング手法に関しては、対象

が検索結果のランキングであるが、ランキング結果として集合を提示する点は我々も考慮する必要がある。情報推薦と比較して、チャート型ランキングが主に提示しているコンテンツ数は少なくないが、普及度を利用したチャート型ランキングにおいては、ユーザの嗜好に対する適合率が著しく低い。そのため、コンテンツ集合の提示を行うことにより、ユーザの嗜好に「サッカー」が含まれるのか、または「なでしこ」が含まれるのか等、類似した特徴語からの確かな嗜好を取得することが出来る可能性がある。

投稿型サイトに投稿されるコンテンツと同様に、ユーザにとって未知のコンテンツであるものとして、ニュース記事が挙げられる。馬ら[8]は、現在の情報フィルタリングでは困難である、ニュース記事のフィルタリングを行った。彼らは過去の配信履歴における時系列的特徴量として新鮮度、流行度、配信頻度を用いた。我々の目標はコンテンツのフィルタリングではないが、チャート型ランキングには類似したコンテンツが同時に上位にランキングされることが多々あり、ユーザの満足度向上や、多様なコンテンツの提示による正確な嗜好の取得のため、同様な考え方を導入を考慮する必要があると考えている。



図6: プロトタイプシステムの実行画面

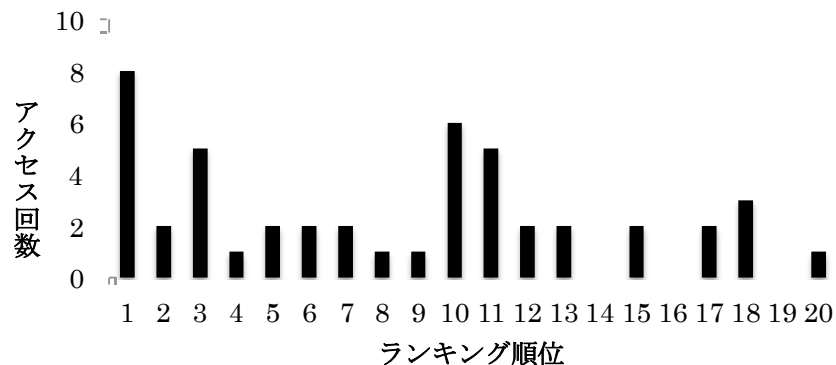


図 7: プロトタイプにより収集されたデータを用いた, ランキング順位とアクセス回数の関係

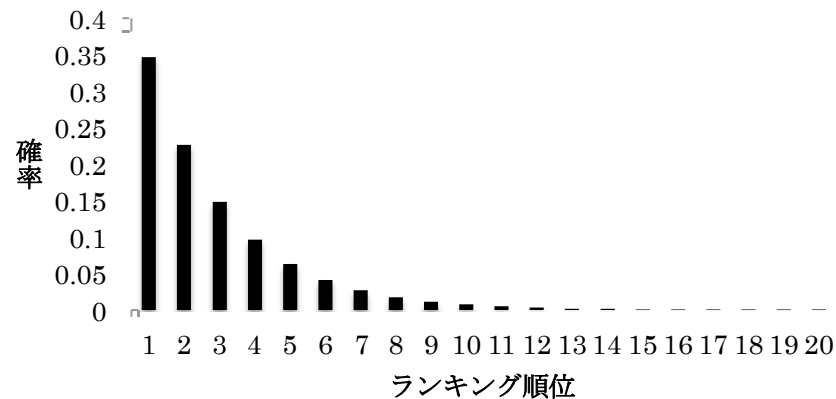


図 8: $\lambda=0.425531915$ のときの $P(r | \text{アクセス})$ の分布

6. まとめ

本研究では, 投稿型サイトにおける適切なコンテンツ提示を行うことを目的として, チャート型ランキングの個人化手法を提案した. チャート型ランキングは, 新規コンテンツが大量に投稿される投稿型サイトにおいても, 未知のコンテンツ提示が可能であるという点で優れている. しかし, 個人の嗜好を全く考慮していないために, ユーザにとって興味の無いコンテンツをも提示してしまう. そこで, 我々はユーザの嗜好と, チャート型ランキングが持つ流行という点の2つを組み合わせることにより, 個人化されたチャート型ランキングを導出した.

今後はプロトタイプシステムを利用して, 提案手法の有用性を検証するための評価実験を行う予定である. 評価実験は, 被験者に対して本プロトタイプシステムをを一定期間にわたって利用してもらい, 同期間のチャート型ランキングとの比較する. また, その他, プロトタイプを利用してもらった際に発生したログをもとに制作したユーザプロフィールに対して, それぞれ取得された特徴語を提示し, 評価を行う等の実験も必要であると考えている.

参考文献

- 1) Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl, GroupLens: an open architecture for collaborative filtering of netnews, Proceedings of the 1994 ACM conference on Computer supported cooperative work, p.175-186, October 22-26, 1994, Chapel Hill, North Carolina, United States
- 2) <http://www.pixiv.net/>
- 3) <http://www.youtube.com/>
- 4) <http://www.nicovideo.jp/>
- 5) 山本 岳洋, 中村 聡史, 田中 克己: RerankEverything: ランキング結果閲覧のための柔軟な再ランキングインタフェース情報処理学会論文誌 (トランザクション) データベース, Vol.3, No.4 (TOD48), pp.48-64, 2010 年 12 月.
- 6) 小林正朋, 五十嵐健夫: Information Fishing: 即応的な情報検索と持続的な情報提示の統合, 第 13 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS 2005) 論文集, 香川, 2005 年 12 月, pp.63-68.
- 7) 湯本 高行, 田中 克己: Web ページ集合を解とするランキング手法, 日本データベース学会 letters 5(4), 13-16, 2007-03
- 8) 馬 強, 角谷 和俊, 田中 克己: 放送型情報配信システムのための時系列性を考慮した情報フィルタリング, 情報処理学会論文誌. データベース 41(SIG_6(TOD_7)), 46-57, 2000-10-15