

Web 閲覧履歴を用いた TV 番組推薦システム

山口 瑤子[†] 瀬々 潤[†]

[†] お茶の水女子大学大学院人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1

E-mail: [†]yamaguchi@sel.is.ocha.ac.jp, ^{††}sesejun@is.ocha.ac.jp

あらまし TV の多チャンネル化に伴い番組を見落としがちになっている。この問題に対応するため、TV 番組推奨システムが提案されている。現在の TV 番組推薦システムは大きく分けて 2 種類あり、1 つは事前に情報登録をするもの、もう 1 つは TV 視聴履歴を利用するものである。前者は利用者に能動的な登録を求めるため負担が大きい。後者は視聴した TV 番組しか利用しないため、推薦できる番組に幅が少ない。これらの問題点を解決するため、本研究では、近年情報収集するための手段としてインターネットの活用が増えていることに着目し、ユーザが検索で用いた検索キーワード、閲覧した Web ページからユーザの多様な興味や趣味を抽出し番組推薦に利用した。また、推薦番組をブラウザ上に表示することで、推薦結果を即時確認し、番組視聴へとスムーズに行動移行ができるシステムを構築した。また、本推薦システムを利用する事で利用者の興味に即した推薦が可能であることを実証する。

キーワード TV 番組, 推薦システム, Web 閲覧履歴

Developing TV Program Recommendation System using personal browsing histories

Yoko YAMAGUCHI[†] and Jun SESE[†]

[†] Graduate School of Humanities and Sciences, Ochanomizu University Otuka 2-1-1, Bunkyo-ku, Tokyo, 112-8610 Japan

E-mail: [†]yamaguchi@sel.is.ocha.ac.jp, ^{††}sesejun@is.ocha.ac.jp

Abstract Recently many TV channels has been provided, and we frequently miss TV programs. Avoiding the missing TV, TV program recommendation systems has been studied. There are two different types of the systems; one requires advance entry about viewers' interest, and the other uses viewing histories of TV programmes. The problems of the studies are requirement of viewer's active operation and limitation of viewer's interest information, respectively. To overcome these problems, our study uses search keywords and Web usage history when viewers use internet because most of the viewers uses internet and the searching behaviour reflect viewers interest. Our system also show our recommendation result on Web browser, which allow us to instantly check TV programs. We show that our system can recommend TV programs which is suitable for viewer interest.

Key words Recommendation System, Browsing History

1. ま え が き

近年 BS デジタル放送や CS デジタル放送が開始され多チャンネル化が進んでいる。これに伴い利用者が閲覧できる TV 番組の種類が増大し、全ての TV 番組の内容を把握することが困難となった結果、変則的に放送される特番や出演者が毎回異なる音楽番組、番組内容が不規則なバラエティ番組など習慣的に視聴しないながら興味ある TV 番組を見逃してしまうことが多くなった。この問題に対し、TV 番組の中から利用者の嗜好に合った TV 番組を表示してくれる推薦システムの研究がされ

ている [1], [2]。電子番組表インターネット TV ガイド [1] では、予め好きな TV 番組のジャンルを登録すると、そのジャンルの TV 番組を紹介する。しかし、利用者がジャンルを予め登録しなければならず、多様な興味を持つ利用者に対しては対応が難しい。この点を改善する手法として、[3], [4] では、視聴履歴から視聴した番組に頻出する出演者を解析し、その出演者が出演する TV 番組を推薦するシステムを提案している。また、視聴者からの番組評価を得て修正するシステムも研究されている [5]。これらの手法は、TV が最も先進的な情報源であった時代には有効に働いていたが、近年は TV 以上にインターネットによる

情報収集が増え、TV は副次的な情報源となっている。TV 番組を内容的なつながりから番組をマップ化し、利用者が番組を辿れるようにする研究 [6] もあるが利用者の能動的な動作が必要なため、日頃の検索結果を利用できる本提案の方が TV の視聴に即していると考えられる。

2. TV 番組推薦システム

本システムは利用者が日頃利用するブラウザから情報を取得し、TV 番組を推薦する。また、利用者が推薦結果を即時確認し TV 番組視聴へとスムーズに行動移行ができるよう、ブラウザ上に TV 番組推薦結果を表示する。

提案システムの概要を図 1 に示す。本システムは、ブラウザ部とサーバ部から構成され、ブラウザ部では利用者のキーワード取得及び推薦番組の表示を行い、サーバ部では TV 番組情報の取得、プロフィールデータの作成、推薦番組の抽出を行う。

2.1 ブラウザ部

ブラウザ部には 2 つの役割がある。ひとつは番組推薦の入力となる検索キーワードを取得しサーバ部に提供すること、もうひとつはサーバ部が推薦した番組の結果を利用者に提示することである。

本研究では両者ともブラウザ Firefox 上で動くプラグインである Greasemonkey を利用して表示する HTML/JavaScript を動的に改変し、サーバへの情報提供、取得及び表示を行った。利用者が利用する検索エンジンとしては、現時点では Google を想定し実装を行ったが Yahoo! や Bing でも同様の仕組みで実装が可能である。利用者の行動から取得するキーワードに関しては推薦システムの実装と評価に関わることであるので、次節で詳細を説明する。

2.2 サーバ部

サーバ部では、利用者が検索エンジンで検索を行った履歴を基に推薦する番組を決定する。

利用者に負担のかからないシステムを目指すため、インターネット利用者が利用する検索窓に入力される単語を利用した。この入力単語を基に番組を推薦することで、利用者が興味ある番組を推薦する事を可能にする。しかし、検索窓に入力された単語から推薦できるテレビ番組は、その単語 (群) に直接的に関連した番組のみであり、その単語を用いて推薦できるテレビ番組は限られる。テレビ番組欄は一般にその番組を凝集して表した言葉であるが、その文、単語数を少なく絞ってあるため、検索語にひっかかるのが難しい場合もあり得た。

本研究では、この問題に対して 2 種類の異なるアプローチを取る。一つは長時間にわたる入力語情報を利用すること、もう一つは検索語の拡張である。前者は、検索履歴を DB に保存しておくことで、過去の検索語を含めて TV 番組推薦を可能にする。後者は、検索後閲覧したページ情報を利用することで検索語の拡張を行う。詳細を以下の節で示す。

本システムの実装では Ruby on Rails を使用し、DB は MySQL を使用している。TV 番組情報は電子番組表 infoseek [7] より取得している。

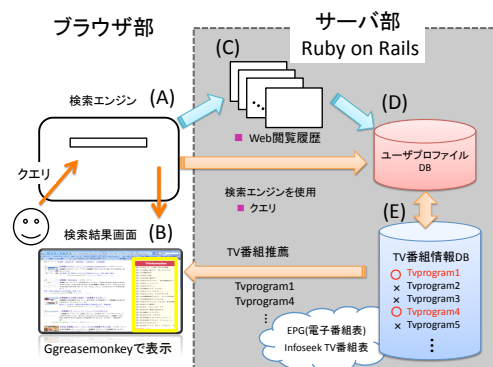


図 1 システム概要図

2.3 検索履歴の作成

検索単語を保存することで、1 回のクエリではなく長期間にわたる利用者の興味を把握できる。このため、検索した単語に時間情報を付与し保存することで、単語の検索時期を記録した。表 1 に例を示す。検索語は単語毎に区切り、別々のレコードとして保存している。また、本テーブルの特徴として、同一単語であっても日付が異なると別のレコードとして記録されている点がある。これにより、その単語に利用者がいつ興味を持ったか把握できる。

表 1 検索履歴

ID	Keyword	Count	Date
1	CookPad	2	2009-12-14
⋮	⋮	⋮	⋮
16	CookPad	1	2009-12-17
⋮	⋮	⋮	⋮
18	CookPad	2	2009-12-19
19	イタリア	9	2009-12-19
20	フィレンツェ	8	2009-12-19
21	美術館	6	2009-12-19
22	CookPad	2	2009-12-20

単純な検索履歴では記録できる単語の数が少なく、テレビ番組を推薦するに足る情報を得られないことが多い。また、検索語は利用者の興味の一部しか表されていない事も多い。たとえば、「イタリア」というクエリに対しては、イタリア料理に興味があるかイタリアの美術館に興味があるかが不明である。そこで、利用者が検索後に移動したページの情報も取得し、検索語に加えることで、両者の問題点を解決した。

遷移後のページから検索履歴に追加する単語は、ページ全ての単語ではなく、ページからそのページに重要な単語を選択して追加した。本研究では、単語の重要度を測る指標として頻繁に利用される tf-idf 法を用いて全単語の重要度を計算し、上位 5 件について検索履歴に追加した。

本拡張により、例えば「イタリア フィレンツェ 美術館」という検索クエリに対して、「ウフィツィ」(美術館)、「ヴィーナス」といった関連のある単語を検索履歴に追加可能である。この追加により「ウフィツィ美術館」が番組情報に含まれるにもかかわらず、「イタリア」が書かれていない場合であっても、推薦結果に含まれる可能性が出てくる。

このシステムでは、追加する単語を得るのに形態素解析ツール ChaSen [8] を利用した。また、tf-idf 法で必要な単語の入ったドキュメント数は anchor、div、paragraph タグで挟まれた文を 1 ドキュメントとしドキュメント数を計算している。

2.4 検索履歴の利用

検索履歴は利用者の嗜好が反映されている単語の集合である。しかし、検索する全ての単語や閲覧した全てのページに対して興味があるとは限らない。そのため、蓄積された検索履歴から期間内に一定比率以上登る単語のみを抽出し、番組の推薦に利用した。

また、推薦に利用する履歴として短期履歴と長期履歴の2種類を表1から用意する。たとえば旅行前の数日間、あるいは、ニュースや広告で見聞きした事による短期的に生まれる興味に対する推薦は短期間の検索履歴を利用し、料理などの趣味の様に長期間に渡って継続する利用者の興味は、長期間の検索履歴を利用する。

たとえば、短期的興味として2日間、長期的興味として7日間を設定し、表1について考える。現在を、12月20日とすると、表1の12月19日、12月20日に蓄積された検索履歴を見る。この2日間に蓄積された検索語の回数の合計は27回であり、10%は2.7回となるので、2.7回以上の回数を持つ検索語を利用する。ここでは、IDが19,20,21のレコードに含まれるイタリア、フィレンツェ、美術館が利用者の短期的興味対象を表す用語とし、推薦番組抽出アルゴリズムに利用する。

長期的興味に関しても同様に抽出するが、利用する期間と単語の数え方が異なる。利用者の長期間に渡る興味は、特定の日に偏って検索するわけではなく、長期間にわたりコンスタントに検索を行うと考えられる。そこで、対象期間の半数以上の日数検索したキーワードを対象として推薦する。表1の12月14日～12月20日の1週間を対象とすると、3回以上出現する単語を対象とする。本例では、CookPadが対象となり、推薦に利用する。

以上、2つの観点から利用者の趣味・嗜好を反映する単語をフィルタし番組を推薦する。

2.5 TV番組情報の処理

本節では、TV番組表の取得と番組推薦に向けた前処理を行う。図1(E)に相当するテーブルの作成である。

TV番組表として、インターネット上のテレビ番組表をパースして利用する事で、利用者がTV番組表を更新することなく、自動的にサーバが必要情報を取得できるようにしている。本研究では、Infoseek電子番組表[7]から各TV番組の情報を入手した。推薦には放送日時、出演者、番組概要を利用した。

表2 TV番組例

タイトル	世界の車窓から
放送日時	2009-12-23 23:10～
出演者	石丸謙二郎
番組概要	イタリア半島南部、シチリア島、サルデーニャ島と6つの鉄道を乗り継いでまわる南イタリア周遊の旅。 本土では、古代都市ボンベイ、洞窟住居の街マテラ、おとぎの国アルペロベッロなど世界遺産の街を訪れます

前処理として、表2から重要語を抽出する。出演者は処理せず推薦に利用する。番組概要はキーワードとなる単語を抽出して推薦に利用する。抽出した例を表3に示す。番組概要からのキーワード選択に

表3

ID	重要語
1	石丸謙二郎
2	シチリア島
4	サルデーニャ島
5	古代都市ボンベイ
6	マテラ
7	アルペロベッロ

はYahoo!のテキスト解析API[9]を用いた。表2の番組概要において、赤くなった部分が取得されたキーフレーズであり、表3のように保存される。このように各TV番組に対し、表3のようなキーワード情報を求め、に保存される。

2.6 推薦番組の抽出

本節では2.4節で求めた利用者の検索語と2.5節で求めたTV番組情報間の相関を調べる事で推薦する番組を決定する。計算したい相関は、利用者の検索履歴とTV番組の関係であるが、本研究では、検索履歴とTV番組から抽出した重要語の相関を求め、その相関を用いて推薦する番組を決定する。

利用者の検索語を x 、TV番組の重要語を y とし、語 w による検索結果のページ集合を $P(w)$ とする。本研究では、語間の相関として高速に計算できるシンプソン係数を利用する。単語 x, y 間のシンプソン係数 $simpson(x, y)$ は、

$$simpson(x, y) = \frac{|P(x) \cap P(y)|}{\min(|P(x)|, |P(y)|)}$$

で定義される。シンプソン係数は値が大きいくほど、単語間の関連が深いことを示している。単語 w が現れるページ数を得るために、ここではYahoo!検索エンジンによる検索結果ページ数を利用した。

例えば、 x =イタリア、 y =フィレンツェとして x と y のシンプソン係数を求めると、 $P(x) = 28,400,000$ 、 $P(y) = 1,770,000$ 、 $P(x) \cap P(y) = 1,560,000$ となり、シンプソン係数は次のように計算できる。

$$simpson(x, y) = \frac{1,560,000}{\min(28,400,000, 1,770,000)} = 0.88$$

このシンプソン係数を利用して、各TV番組の得点を計算する。利用者の短期履歴、長期履歴から抽出された各検索キーワードとその各検索キーワードによって得られたWebページの中から利用者が閲覧したWebページの重要単語群を $X = \{x_1, \dots, x_n\}$ 、あるTV番組から抽出した重要語群を $Y = \{y_1, \dots, y_m\}$ とする。以下の指標で、利用者とTV番組の関連得点を計算する。

$$point(X, Y) = \frac{1}{nm} \sum_{x \in X, y \in Y} simpson(x, y)$$

この得点が高いTV番組上位10個を推薦番組としてWeb画面に表示する。

3. 実験

本研究では、検索語の拡張、及び、履歴を短期、長期に分けて利用する手法を導入したが、これらの手法の有用性を検証する。

3.1 検索履歴拡張の有用性

利用者が閲覧したWebページから抽出した重要単語を推薦に追加することの有用性についての検証を行った。 X_s を検索語のみを要素とする検索履歴単語、 X を検索語に加え閲覧されたWebページから抽出した重要単語を要素とした履歴とし、一週間分のTV番組に対して、得点の高かった番組の検証を行った。

ここでは、利用者がネットで配信されているニュース

表 4 推薦番組の比較

順位	X _s による推薦番組	X による推薦番組
1	NEWS 23	聖志のゴルフ宮里流
2	報道ステーション	村口史子のグッドゴルフ
3	NHK杯囲碁トーナメント	京セラ フェニックスチャレンジ
4	芸能界の告白特別編	産経テレニュースFNN
5	スジナシ	土曜スポーツタイム
6	スーパーJチャンネル	牧野裕のEnjoy Golf
7	N響アワー	NHK週刊ニュース
8	FNNニュース	NNNニュース
9	アナ★パン!	ゴルフの真髄
10	囲碁の時間	Get Sports

を閲覧することを想定する。検索キーワード $X_s = \{\text{yahoo ニュース}\}$ 、で検索し、表示されたニュースの中からゴルフに関する Web ページ[10]を閲覧した結果、 $X = \{\text{yahoo ニュース、石川、宮里、池田、産経新聞}\}$ が得られた。表 4 は X_s に基づいて推薦された TV 番組と X に基づいて推薦された TV 番組、それぞれ上位 10 番組を比較している。 X による推薦番組において、順位=1,2,3,6,9,10 の番組はゴルフに関する内容の番組である。上位 10 番組においてゴルフに関する番組は、 X_s の場合 0%、 X の場合 60% となった。このように、検索キーワードだけではなく、利用者が閲覧した Web ページを追跡し重要単語を抽出することで、ゴルフという利用者の趣味を特定し、番組推薦の結果に反映することができた。

3.2 短期と長期に履歴を分ける事の有益性

表 5 履歴の期間の影響

	料理番組	ニュース・特集番組	その他
短期	26%	13%	61%
長期	2%	61%	27%
両方	13%	47%	40%

本手法では検索履歴に用いた単語を現れた単語の回数を利用する短期、単語の日数を利用する長期に分けて抽出していた。本節では、この短期及び長期に分けたことの影響を検証する。

表 5 は短期履歴から選ばれた単語を「インフルエンザ」、長期履歴からの単語として「CookPad」とした場合の、推薦された番組の各ジャンルに対する割合を表している。表 5 から、短期履歴では料理番組を取得することができず、長期履歴ではインフルエンザが含まれるであろうニュース・特集番組を取得することができないことが分かる。一方、これらの情報を複合して推薦に用いることで、料理番組、ニュース・特集番組共に取得することができている。このように、複数の観点から抽出された単語を推薦に用いることで利用者の興味に対し、幅広い番組を抽出することができる。

3.3 推薦番組の検証

表 6 推薦番組例

順位	推薦番組
1	龍馬伝×プロフェッショナル
2	首都圏ネットワーク
3	爆笑問題のもうひとつの龍馬伝
4	ニュースウオッチ9
5	NHKニュース おはよう日本
6	龍馬伝
7	J ブンガク
8	EXILE GENERATION
9	地球の目撃者SP
10	ビタゴラスイッチ

実際に得られた検索履歴を基に推薦番組を抽出し、検索履歴の要素に関連した TV 番組が推薦されているか検証を行った。 X を検索語に加え閲覧された Web ページから抽出した重要単語を要素とした履歴とする。ここでは、利用者が大河ドラマの内容について調べる為に検索キーワード「大河ドラマ」で検索し、Web ページ：大河ドラマ「龍馬伝」[11]を閲覧した結果、履歴情報 $X = \{\text{大河ドラマ、福山、広末、龍馬、長崎}\}$ が得られた。表 6 は、この X に基づいて推薦された TV 番組である。順位=1,...,6 の TV 番組いずれにおいても、TV 番組の情報として「福山雅治」が含まれており、 X の要素である「大河ドラマ」、「福山」、「龍馬」と高い相関を示している。また、順位=1 の TV 番組では、TV 番組情報として「龍馬×プロ」が含まれており、 X の要素「龍馬」と高い相関を示している。順位=2 の TV 番組では、TV 番組要素として「福山雅治」以外に、大河ドラマ「龍馬伝」の出演者である「香川照之」、「池田達郎」が含まれており、 X の要素である「大河ドラマ」、「龍馬」とそれぞれ高い相関を示している。順位=4,5 の TV 番組では、TV 番組要素として「龍馬伝」が含まれており、 X の要素「大河ドラマ」、「福山」、「龍馬」と高い相関を示している。このように、検索履歴の要素に関連した TV 番組を抽出することができた。

4. ま と め

本研究では利用者の検索履歴と閲覧ページを利用することで、能動的な動作無く興味ある TV 番組を推薦するシステムを作成した。また、本システムを利用し、最近のニュースからゴルフに関するページを閲覧する事で、ゴルフ番組を推薦する事や、大河ドラマから指定の年の大河ドラマのページを見ることで、関連番組が推薦される事を検証した。TV 番組欄には情報が少ないため、単純な推薦では「情報」などの一般的な用語が上位にランクされやすい点を回避し、ユーザの興味に即したページを抽出できたことを示している。今後の課題として、Web ページの解析手法を改善し、システムの推薦精度を高めたい。

文 献

- [1] TV ガイド: <http://www.tvguide.or.jp/>
- [2] Yahoo!テレビ.G ガイド: <http://tv.yahoo.co.jp/>
- [3] 土屋誠司、佐竹純二、近間正樹、上田博唯、大倉計美、蚊野浩、安田昌司、“TV 番組推薦システムの構築とその有用性.” 情報処理学会 研究報告 pp.95-102 2006011
- [4] 宮原浩二、小谷亮、小川吉大、小林啓二、近藤省造、“利用者の視聴履歴に基づく TV 番組推薦システムの検討,” 情報処理学会 第 54 回全国大会平成 9 年前期 (4) pp.245-246 19970312
- [5] Kaushal Kurapati, Srinivas Gutta, David Schaffer, Jacquelyn Martino, John Zimmerman A Multi-Agent TV Recommender In Proceedings of the UM 2001 workshop
- [6] 溝口祐美子, “オントロジーを用いた TV 番組グラフ作成システム.” 電子情報通信学会 WI2 第 8 回研究会
- [7] Infoseek 電子番組表: <http://tv.infoseek.co.jp/>
- [8] ChaSen: <http://chasen-legacy.sourceforge.jp/>
- [9] Yahoo!キーワード抽出 API: <http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html>
- [10] yahoo ニュース: <http://headlines.yahoo.co.jp/hl?a=20100101-00000540-san-golf>
- [11] 大河ドラマ「龍馬伝」: <http://www9.nhk.or.jp/ryomaden/>