Web 閲覧履歴を用いた TV 番組推薦システム

人間文化創成科学研究科 理学専攻 情報科学コース　山口 瑶子　(指導教員：瀬々 潤)

1　はじめに

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。これに伴いユーザが閲覧できる TV 番組の種類が増大し、全ての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など習慣的に視聴しないながら興味ある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 番組の中からユーザの嗜好に合った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。しかし、どのシステムにおいてもユーザの嗜好を抽出する情報源が TV 番組に関する情報に限られており、ユーザの多様な興味を反映した番組を推薦することができない。そこで、本研究では多くのユーザがネットや携帯を見ながら TV を視聴することに着目し、Web の閲覧履歴から得られるユーザの幅広い興味や趣味を利用し、TV 番組を推薦するシステムを構築した。

2　関連研究

現在 TV 番組推薦システムは数多く存在している。インターネット TV ガイド [1]では、予め好きな番組のジャンルを登録すると、そのジャンルの番組を紹介する。しかし、ユーザがジャンルを予め登録しなければならず、多様な興味を持つユーザに対しては対応が難しい。この点を改善する手法として、[1, 2] 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番組を推薦するシステムを提案している。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。これらの社会を取り巻く状況の変化に対応し、本研究ではインターネットで検索した用語及び閲覧履歴を利用する事で、インターネットに連動した TV 番組情報取得を可能にする。

3　TV 番組推薦システム

本システムはユーザが日頃利用するブラウザから情報を取得し、ユーザ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、ブラウザ上に推薦結果を表示する。提案システムの概要を図 1 に示す。

3.1　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事、もうひとつはサーバ部が推薦した番組の結果をユーザに提示することである。

3.2　サーバ部

サーバ部では、ユーザ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。ユーザに負担のかからないシステムを目指すため、インターネットユーザ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、ユーザが興味ある番組を推薦する事を可能にする。

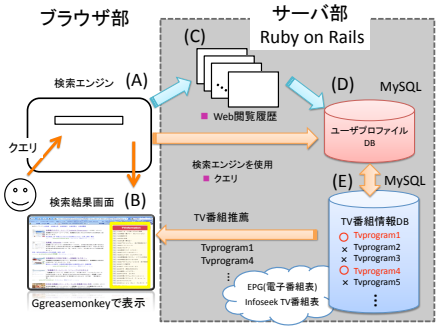

図 1: システム概要図

3.3　検索履歴の作成

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する。また、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することでその単語にユーザがいつ興味を持ったか把握できる。

表 1: プロファイルデータ

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、検索語はユーザの興味の一部しか表されていない事も多い。そこで、ユーザ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。ページからの単語抽出には形態素解析ツール ChaSen[2]を利用している

3.4　検索履歴の利用

ユーザにとって検索すること全てに対し興味があるとは限らない為、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、

[1]http://www.tvguide.or.jp/

[2]http://chasen-legacy.sourceforge.jp/

12 月 20 日とすると、表 1 の 12 月 19 日、12 月 20 日に蓄積されたプロファイルデータを見る。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、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。

3.5 TV 番組情報の処理

本研究では、Infoseek 電子番組表[3]から各 TV 番組の情報を入手し図 1(E) に相当するテーブルを作成した。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、サルデーニャ島、シチリア島と 6 つの鉄道を乗り継いでまわる南イタリア周遊の旅。

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は Yahoo! のテキスト解析 API[4]を用いた。

3.6 推薦番組の抽出

表 3: 番組概要の重要語

ID	重要語
1	石丸謙二郎
2	シチリア島
4	サルデーニャ島

本節では 3.4 節で求めた利用者の検索語と 3.5 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。

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 ページの中からユーザが閲覧した Web ページの重要単語群を $X = \{x_1, \ldots, x_n\}$、ある TV 番組から抽出した重要語群を $Y = \{y_1, \ldots, y_m\}$ 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。

4 検証

表 4: 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	ピタゴラスイッチ

プロファイルデータの要素に関連した 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、ユーザ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」[5]を閲覧した結果、X = { 大河ドラマ、福山、広末、龍馬、長崎 } が得られた。表 4 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、番組情報として「福山雅治」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。また、順位=1 の番組では、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 の番組では、番組要素として、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「大河ドラマ」、「龍馬」と高い相関を示している。このように、プロファイルデータの要素に関連した TV 番組を抽出することができた。

5 まとめ

本研究ではユーザの Web 閲覧履歴、検索キーワードから、ユーザの幅広い興味や趣味を抽出し番組を推薦するシステムを構築した。今後、Web ページ解析手法を改善しシステムの推薦精度を高めたい。

参考文献

[1] 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 20060113

[2] 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

[3] Kaushal Kurapati, Srinivas Gutta, David Schaffer, Jacquelyn Martino, John Zimmerman A Multi-Agent TV Recommender In Proceedings of the UM 2001 workshop

[3] http://tv.infoseek.co.jp/

[4] http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html

[5] http://www9.nhk.or.jp/ryomaden/