

TV 番組録画システムと好み自動学習

○真船 勉, 杉山 雅英 (会津大学)

1 まえがき

現在, 家電量販店などには多くの HD, DVD レコーダーが並んでいる。これらの製品は使いやすく優れた機能を持ったものが多い。しかし製品の内部仕様, アルゴリズムなどは公開されておらず, 新しい機能をユーザーが自分で追加することは困難である。

本報告は Linux 上で TV 番組を視聴, 録画するシステムの構築を目標としている。コンピュータ上でシステムを実現することでユーザーが自分で機能の追加, 改良が出来るようになり, 市販の HD, DVD レコーダーの問題点を改善することができる。

さらに本報告では新しくこのシステムに追加した「個人の好みの学習による自動録画予約機能」について報告する。ユーザーの興味に適した情報を推薦, フィルタリングするユーザープロファイリング技術が述べられている [1]。市販の製品にもユーザーの好みを学習して自動録画をする機能を持った製品 [2] があるが, そのアルゴリズムは公開されておらず, ユーザーが機能を改良したり, 新たな学習機能を追加することはできない。

2 OpenPVS

devicehackers.com 社開発の OpenPVS とは既存の Linux ディストリビューションをベースとし, フリーソフトウェア群, 一般的に入手可能なハードウェアで構成されたパーソナルビデオサーバーである [3]。本報告ではこの OpenPVS をベースとし, 新しい機能を追加する。以下では OpenPVS の重要な構成ソフトウェアである MythTV について述べる。

OpenPVS はビデオレコーダーソフトである MythTV を用いて TV 録画機能を実現している [4]。また MythTV には TV 視聴, 録画機能以外にも例えば, 音楽再生アプリである MythMusic など多くのプラグインが用意されており必要に応じてこれらを組み込むことにより, 機能を拡張することができる [5]。MythTV はフロントエンドプログラム (mythfrontend) とバックエンドプログラム (mythbackend) に分かれている。TV 番組録画などのメイン処理はバックエンドプログラムが行い, フロントエンドプログラムはユーザーとのやりとりを行う。この仕組みの利点としては, バックエンドプログラムとフロントエンドプログラムを異なるコン

ピュータで動作させることが可能な点である。プラグインである MythWeb を導入すれば Web ブラウザをフロントエンドとして使用することも可能である。また MythTV は全てのデータをオープンソースのリレーショナルデータベースソフトウェア MySQL が管理している。これは電子番組表 (EPG) による録画予約において膨大なデータを使用するためである。

3 個人の好みによる自動録画予約

個人の好みによる自動録画予約機能は, 個人の好みを予めシステムに学習させ, EPG より取得した TV 番組データからその TV 番組がユーザーの好みの番組であるかを判別し自動録画する機能である。

3.1 個人の好みの学習

個人の好みの学習はユーザーの一定期間の録画の履歴を用いて行うものであるが, 本報告ではユーザーの好み学習方法として, 過去の番組データを呈示しユーザーに好みの番組を指定させることとした。過去 1 ヶ月分の約 11000 個の番組データからランダムにユーザーに表示し, 各番組に対して好み/非好みを 1-0 で指定してもらう。ユーザーに対して呈示する番組データは番組タイトル, 番組の内容である。

3.2 P2filter

P2filter とはスパムフィルタリングソフトである bsfilter [6] を元に開発した好み判別 filter である。bsfilter はベイズ理論に基づいてユーザーの好みを判別する。本報告では Gary Robinson 法 [7] を用いた。番組データの文字列を単語に分け単語毎に好みである確率を求める処理, そして指定された TV 番組がユーザーの好みであるか判断する処理から成り立っている。bsfilter が番組データに含まれる全ての品詞の単語を使用するのにに対して, P2filter は使用する単語を指定の品詞等に限定する。

P2filter では好みの番組の判定を行う前に単語 w 毎に好みの番組に含まれる確率 $p(w)$ を求める必要がある。単語の好みの番組内での確率を $p_s(w)$, 非好みの番組内での確率を $p_n(w)$, 単語を含む番組が好みの番組だった数を $N_s(w)$, 単語を含む番組が非好みだった数を $N_n(w)$, そして好みの番組総数, 非好みの番組総数をそれぞれ T_s, T_n とすると $p(w)$ は以下の式 (3) で求めることができる。

$$p_s(w) = \frac{N_s(w)}{T_s} \quad (1)$$

*TV Program Recording System and Automatic Learning Personal Preference,
by T.Mabune, M.Sugiyama (The Univ. of Aizu)

$$p_n(w) = \frac{N_n(w)}{T_n} \quad (2)$$

$$p(w) = \frac{p_s(w)}{p_s(w) + p_n(w)} \quad (3)$$

判定する番組内に出現する単語 w の好み確率を $f(w)$ とする. 全単語での $p(w)$ の平均値を m , その単語の学習データ内での出現数を N , そしてある正の定数を c とし $f(w)$ を以下の式 (4) で定義する. なお本報告において $c = 0.001$ とした.

$$f(w) = \frac{c \cdot m + Np(w)}{c + N} \quad (4)$$

もし $N = 0$ であれば, $f(w) = m$ となる.

番組の好み確率はその番組に含まれる全ての単語の $f(w)$ の結合確率によって定義する. S を番組の好み確率とすると, S は P と Q より以下の式 (5) で求められる.

$$\begin{cases} P = 1 - (\prod_{i=1}^I (1 - f(w_i)))^{\frac{1}{I}} \\ Q = 1 - (\prod_{i=1}^I f(w_i))^{\frac{1}{I}} \end{cases}$$

$$S = \frac{P}{P + Q} \quad (5)$$

$Q = 0$ 即ち $\forall i, f(w_i) = 1$ のとき $S = 1$ となる. I は判別する番組データ内の単語数である.

3.3 P2filter の評価実験

被験者に 1000 タイトルの TV 番組に対して好みを指定させ 700 番組で学習, 300 番組で評価を行った. 実験者数は男 5 人, 女 2 人の計 7 名で全て学生である. 使用する単語の集合を以下の 2 つのものとして実験をおこなった. (1) 空白, 記号, 助詞を除く, (2) 名詞のみ

集合 (1) に対する実験結果を検証すると, 使用する単語の品詞が名詞のものであるものが重要であると推測された. 集合 (2) に対する実験結果を図 1 に示す. 横軸は番組の好み確率 S , 縦軸はその確率である番組が好み/非好みそれぞれの総番組数に占める割合を表す. 実線はユーザーが好みと指定した番組の確率, 破線は好みではないと指定した番組の確率を示している. 実線と破線の分布が重ならなければ, ある閾値で好みの番組を判別出来ることを表している. 好みの番組であるのに閾値を下回るものの割合を好みエラー率, 非好みの番組であるのに閾値を上回るものの割合を非好みエラー率と定義し, それらの平均値を平均エラー率と呼ぶ. 図 1 の実験結果において平均エラー率が最小になるように閾値を 0.42 と設定すると, 好みエラー率は 30.7%, 非好みエラー率は 12.6% となった. 集合 (1) に対する実験結果と比較して集合 (2) に対する実験結果は平均エラー率が 1.5% 減少した. なお個人毎の最適な閾値の最大値は 0.51, 最小値は 0.29 となり, 平均エラー率は最大 35.2%, 最小 10.2% となった. 個人によって最適な

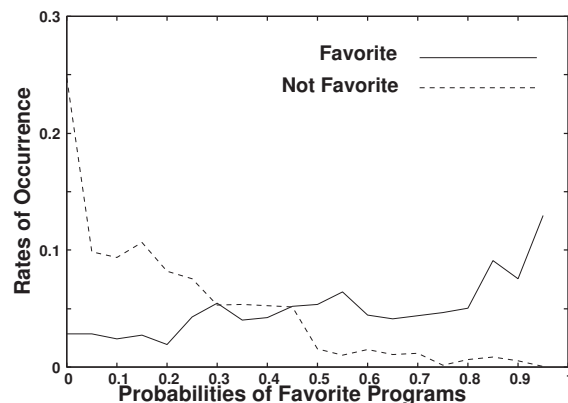


図 1: P2filter を用いた好み学習の評価実験結果

閾値が大きく離れることが分かる. 最大エラー率が大きくなった要因としては学習, 評価用のサンプル数が少なかったことが考えられる.

4 むすび

本報告では Linux 上で TV 番組を視聴, 録画するシステムを構築し個人の好みによる自動録画予約機能を新しい機能として追加した. また自動録画予約機能に使われる好み判別フィルターである P2filter の評価実験を行った. 番組データに含まれる単語のうち名詞のみを使用すると好みの番組を判別することが出来た.

今後の課題としては OpenPVS への新しい機能の追加, P2filter の性能向上が考えられる. ユーザーごとに最適な閾値を設定する機能も必要である.

謝辞

研究を進める上でご支援頂いたヒューマンインターフェース学講座の皆様, 好みの学習に協力してくれた皆様に感謝いたします.

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ープロファイリング技術”, 人工知能学会誌, vol.19, no.3, pp365-372, 2004
- [2] SONY, “CoCoon”, <http://www.sony.jp/products/Consumer/cocoon>
- [3] devicehackers.com, “OpenPVS”, <http://www.devicehackers.com/openpvs/index.php>
- [4] mythtv.org, “MythTV”, <http://www.mythtv.org/index.php>
- [5] 伊藤ぼん太, 吉川敦, “MythTV 活用ガイド”, (株) 九天社
- [6] nabeken, “bsfilter”, <http://bsfilter.org>
- [7] Gary Robinson, “Spam Detection”, <http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html>