F-029

ベイジアンフィルタを利用した Web 推薦システム
Web Recommendation System Using Bayesian Filter

庭野 正義† マッキン ケネス ジェームス‡ 永井 保夫‡
Masayoshi Niwano Kenneth James Mackin Yasuo Nagai

1. はじめに

Google などに代表される検索エンジンを用いて Web ページを検索する場合、検索結果は膨大となり、必ずしもユーザ個人に最適化されているとは限らない。大量の検索結果の中から必要なページを判断するにはかなりの労力が必要となる。Web ページ推薦の手法として、協調フィルタリングの手法で推薦を行う研究がなされている[1]。しかしながら、一般的にはユーザが少ない場合にうまく推薦ができない、Web ページの数が膨大で、共通したアイテムを持っているユーザが少ないためうまく推薦ができない、一部の Web ページのみを対象とした推薦に留まっている、などの問題点が指摘されている。

本研究では、このような問題点を解決することを目標に、ベイジアンフィルタを利用した Web 推薦システムを提案する。

2. 提案する Web 推薦システムの構成

2.1 提案するシステムの特徴

Web ページを検索する研究では、ベイジアンフィルタを用いる手法がすでに提案されており[2]、類似性の高い Web ページを収集することに成功している。しかしながら、この手法ではユーザが評価結果をいちいちシステムに入力しなければならないという手間が生じる。

本研究では、ユーザの評価情報入力の負担を減らすために、國貞らの研究[3]のように、既存の検索システムから返された検索結果(タイトル、概要、ホスト名)をユーザに評価してもらい、評価結果に基づいて推薦を行う手法を検討した。この手法により、検索結果を選択するという行為が、検索結果の評価とシステムへの入力になり、システムに学習させるという手間を無くすことができる。システム構成を図 1 に示す。

2.2 学習および推薦の手順

1) 検索語の受け取り

一般的な検索システムと同じように、ユーザが検索したい語を入力すると、制御部がそれを受け取る。

2) 検索語での検索

制御部は、受け取った検索語を検索システムに送信し、検索結果を受け取る。

今回は Google Ajax Web API を使用し、検索結果を取得する。今回取得する検索結果の数は、Google Ajax Web API で取得できる最大数の 32 個とした。

3) カテゴリの取得

制御部は、検索語から検索結果のカテゴリを取得する。

例えば、"java"という語句で検索した場合、検索システムから返ってくる検索結果は"java"カテゴリに属しているだろうことが予想される。提案するシステムでは、検索キーワードを形態素解析部に与えて解析を行い、解析結果である形態素を検索結果のカテゴリとした。"ベイズ理論"で検索した場合、検索結果は"ベイズ"と"理論"の 2 つのカテゴリに属することになる。

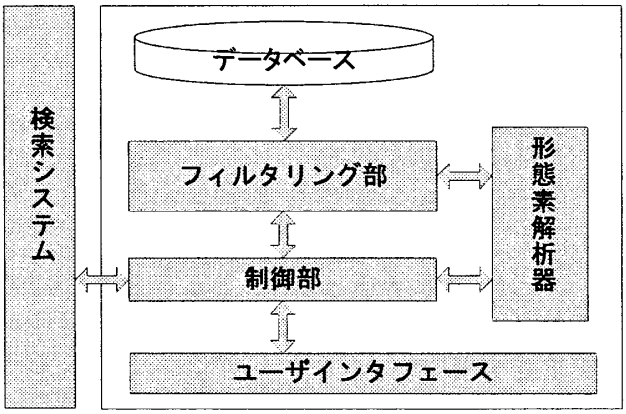

図 1：システム構成

4) 推薦度の計算とユーザへの提示

制御部は、フィルタリング部に検索結果の解析を依頼し、検索結果の順位付けを行い、検索結果を推薦度の降順に並べ替えユーザに提示する。

5) ユーザの判断に基づいた嗜好情報の収集

最終的にユーザがどのアイテムを選択したかという情報は制御部に渡される。制御部は、推薦したのに選ばれなかったアイテムのトークンと、推薦しなかったのに選ばれたアイテムのトークンをフィルタリング部にそれぞれ追加記録させる。

2.3 ユーザ嗜好の記録

ユーザの嗜好情報は、表 1 のようにデータベースに格納される。例えば、ユーザが、"Linux インストール" という検索語で検索した場合、"サーバ"という単語が含まれている文章に興味を持った回数が 1、興味を持たなかった回数が 3 だとする。この場合、データベースには表 1 のように保存される。このデータを基に、カテゴリに属するトークンの推薦度を計算する。

表 1：データベース構成の具体例

word	category	match_count	nonmatch_count
サーバ	Linux	1	3
サーバ	インストール	1	3

2.4 検索結果の順位付け

検索結果 R(Result)が返ってきた時、ユーザが R に興味 I(Interesting)をもつ確率を、ベイズの定理より(1)式で求めることができる。

$$P(I \mid R) = \frac{\prod_{i=1}^{n} P(I \mid T_i)}{\prod_{i=1}^{n} P(I \mid T_i) + \prod_{i=1}^{n} (1 - P(I \mid T_i))} \quad \cdots(1)$$

ここで、P(I|R)は文章中の 番目のトークンにユーザが興味を持つ確率であり、『ユーザが過去にこのトークンが含

†東京情報大学 大学院総合情報学研究科 総合情報学専攻
‡東京情報大学 総合情報学部 情報システム学科

まれている文章に興味を持った回数』を『ユーザが見た文章のうち、このトークンが含まれている文章すべての数』で割ったものとなる。それぞれの回数は、前述のデータベースを参照することで得られる。(1)式を使い、カテゴリ毎に文章の推薦度を求め、最後に全カテゴリの推薦度の結合確率を求める。ここで求めた確率を推薦度とし、検索結果を推薦度の降順に並び替え、ユーザに表示する。

3. 提案した推薦システムを用いた実験と結果

3.1 実験方法

まず、ユーザに Google Ajax Search API を使用して検索をしてもらう。その時の検索キーワード、検索結果の集合、ユーザの評価結果を記録する。

次に、推薦システムにより順番を入れ替えた結果と Google API の検索結果について、上位 4 つの検索結果に入っている適合文章の数を比較する。この時、検索結果の中に適合文章が入っていない場合、検索は無視する。

3.2 実験結果と考察

検索結果の比較を表 2 に示す。表 2 は、"アプリオリアルゴリズム 信頼度"という検索語で検索した場合の結果を比較したものである。過去には"アプリオリアルゴリズム 支持度"、"支持度"、"支持度とは"という検索を行ったことがあり、その学習データを用いて、推薦度を計算した。

表 2：表示順番の比較

	Google API の検索結果	推薦結果
1	*データマイニングアルゴリズム“ アプリオリ ”と“ ”の比較	*Walk This way の開発
2	アソシエーション分析 (association analysis／相関分析) ｜ ○○商会	*データマイニングアルゴリズム“ アプリオリ ”と“ ”の比較
3	データマイニングアルゴリズム「アプリオリ」と「ID3」の比較	アソシエーション分析 (association analysis／相関分析) ｜ ○○商会
4	○○ ○博士（工学）	データマイニングアルゴリズム 「アプリオリ」と「ID3」の比較
5	食品業界向け文字照合センサが 安全性確保・履歴明確化の重要性に応える。	○○ ○博士（工学）
6	*Walk This way の開発	DBMS_DATA_MINING
7	DBMS_DATA_MINING	情報洪水時代における アクティブマイニングの実現 研究成果報告書
8	情報洪水時代における アクティブマイニングの実現 研究成果報告書	食品業界向け文字照合センサが 安全性確保・履歴明確化の重要性に応える。

「*」マークは、マークのついたページにユーザがアクセスしたことを示す。入れ替えにより 1 位となった「Walk This way の開発」というアイテムを解析した結果を表 3 に示す。特徴的なトークンとその推薦度の項目は、検索システムから受け取った「Walk This way の開発」のタイトル、概要、ホスト名をひとまとめにした文章を解析した結果の中から特徴的なトークンを取り出したものである。「カテゴリ：トークン＝ 推薦度」で表されており、「アプリオリ：アルゴリズム＝0.90...」は、アプリオリカテゴリに属するアルゴリズムというトークンの推薦度が 0.90...であるということを示している。本システムにより推薦された「Walk This way の開発」は、パーソナルルート提示システムの開発について書かれている論文である。その中で、アプリオリアルゴリズム、信頼度、支持度についての説明がされている。このように、関連性の高い検索語で繰り返し検索をした場合、有効な推薦ができる可能性が高いことが明らかになった。

表 3：「Walk This way の開発」の解析結果

検索語:"アプリオリアルゴリズム 信頼度"
カテゴリ:"アプリオリ", "アルゴリズム", " 信頼", "度"
特徴的なトークンとその推薦度： アプリオリ：アプリオリ＝0.9 アプリオリ：アルゴリズム＝0.9090909090909091 アプリオリ：度＝0.9473684210526315 アプリオリ：支持＝0.9090909090909091 アルゴリズム：アプリオリ＝0.9 アルゴリズム：アルゴリズム＝0.9375 アルゴリズム：度＝0.9473684210526315 アルゴリズム：支持＝0.9090909090909091 アルゴリズム：jp＝0.9166666666666666 度：確信＝0.9
カテゴリ毎の推薦度： アプリオリ＝0.999938275 アルゴリズム＝0.999996259 信頼＝0.5 度＝0.9
全体の推薦度：0.9999886634168462

4. おわりに

本論文では、ベイジアンフィルタを利用した Web 推薦システムを提案し、3 章の事例のようなケースでは有効な推薦ができることを示した。さらに、本実験では、以下のような問題点についても明らかになった。

- 特徴的なトークンとして、推薦度が 0.9 以上か 0.1 以下のトークンに限定して解析をしたが、閾値の調整を行い検証する必要がある。
- 検索語毎に独立して推薦度を計算しているため、初めての検索語を検索した場合に推薦ができない。

今後は、上記のような問題点を改善しつつ、学習量を増やして実験を行っていきたい。また、協調フィルタリングとの融合についても検討していく予定である。

参考文献

[1] 高須賀清隆, 丸山一貴, 寺田実: 閲覧履歴を利用した協調フィルタリングによる Web ページ推薦とその評価, 電子情報通信学会技術研究報告 Vol.107, No.131, pp. 115-120, 2007 年.

[2] 天野環, 中里秀則, 中村隆史 ：ベイズ推定を用いた Web マイニング, 電子情報通信学会技術研究報告 Vol.104, No.724, pp. 43-48, 2005 年.

[3] 國貞暁, 山本けい子, 田村哲嗣, 速水悟 ：要約情報の類似度を用いた WEB 検索支援システム, 第 21 回人工知能学会全国大会, 2007 年.