

3U-1

情報フィルタリング技術を用いた Web 情報検索の提案

八鍬 健太[†] 辻 秀一[‡]

東海大学大学院工学研究科[†]

東海大学情報メディア学科[‡]

1. はじめに

現在、Web 情報量は 1 億ページを超える。そのような中で、問題視されているのが Web 上から見つけたい「情報」が見つげにくくなっているということである。それにより、Web 情報検索はクエリを 1 語入力しても良い結果が得られなくなってきた。

そこで個人適応化 Web 情報検索システムにより、個人履歴情報から個人の嗜好に合った検索結果を求める提案をした。本研究はその改良版として個人支援型 Web 情報検索システムを提案する。

2. 従来の情報フィルタリング技術

2-1 情報フィルタリングとは

情報フィルタリングとは情報空間から必要な情報を抽出する事である。そして、その情報フィルタリングの種類として協調フィルタリングとコンテンツベースフィルタリングといった 2 種類のものがある。今回は協調フィルタリングを使うので、そちらについて述べようと思う。

協調フィルタリングは他の同じ嗜好を持つユーザの情報を用いて、情報の推薦を行う仕組みで、メモリベースとモデルベースといった種類がある。メモリベースのものとして、ピアソンの相関関数がある。これは 2 変数 X, Y が n 組あるとし、ピアソンの積率相関係数 r は、「変数 X と変数 Y の共分散」と「それぞれの変数の標準偏差」から求められる。また、その方法に使われている最短距離法 (nearest neighbor method) といったものがある。この相関係数の値は $-1 \sim 1$ の間の値をとる。値が 1 に近いほど 2 つのベクトルの相関が強く、 -1 に近いほど逆相関が強い。そして、その絶対値は 0 に近いほど相関関係が無く、1 に近づくほど相関関係が強い。

2-2 現在の Web 情報検索システムの問題点

現在、Web 情報検索システムの一の問題として、同じクエリ (検索語) で検索したユーザはどのユーザが検索しても同じ検索結果が出てしまうといった問題がある。そこで、本研究では個人適応検索システムの中核として情報フィルタリング技術を用いた Web 情報検索システムにより以上の問題点を解消することを目的として、より良い検索システムの実現を目指す。

3. 個人支援型 Web 情報検索システム

3-1 提案システム全体概要

本研究の提案システム全体図を Fig.1 に示す。本システムは情報フィルタリング技術のひとつである協調フィルタリングを使って、Web 情報検索で個人を支援するシステムである。

具体的なシステムの提案としては、まずユーザが Web 情報検索を使って検索を行い、その閲覧履歴を収集し、個人プロフィールを生成する。そして、次に検索を行う場合にユーザはクエリを 1 語入力し、そのクエリに対して関連したもうひとつのクエリを個人プロフィールに基づいて、自動的に追加を行って検索を行う支援システムである。支援システムの主な部分として、フィルタリング部、検索部、個人プロフィール部といった 3 つの部分で構成されている。

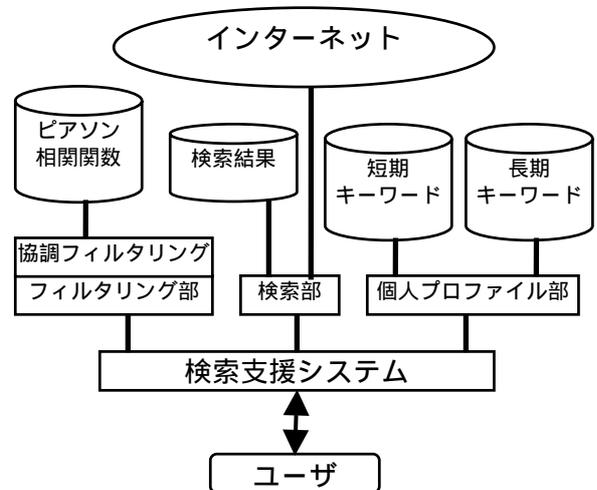


図.1 システムの全体図

Web information searching system that using information filtering

[†]Kenta Yakuwa

Graduate School of Engineering, Tokai University

[‡]Hidekazu Tsuji

School of Information Technology and Electronics, Tokai University

本システムは「検索部」「情報フィルタリング部」「個人プロフィール部」の3つのモジュールで構成される。

3-2 情報フィルタリング部

本部分は「協調フィルタリング部」から成り立つ。まず、協調フィルタリングとはユーザ対グループの評価値行列を作る場合、評価値を予測するのに必要なものである。そして、協調フィルタリングの方法として最短距離法 (nearest neighbor method) が挙げられる。この方法ではまず、対象となるユーザ群に対して、類似度の重み付けを行う。その場合の類似度は評価値のベクトル計算でピアソンの相関係数を使う。簡単な式としては

$$r = \frac{\text{変数 X と変数 Y の共分散}}{\text{変数 X の標準偏差} \times \text{変数 Y の標準偏差}}$$

といった式となる。

そして、対象となったユーザの中から、高類似度をもつ 2~3 人のユーザを同じ嗜好を持ったユーザとする。そして、そのユーザの評価値を使って、その予測値を計算するといったことを行う。この協調フィルタリングを使うことで、検索の「検索の網羅性」と「個人の嗜好」とを両立させた検索が可能になる。

3-3 検索部

本部分では検索エンジンを使い、ユーザが検索結果からページを閲覧する。そして、その閲覧履歴の Web ページからフィルタリング部にデータを送る。そして、次の検索を行う場合には、個人プロフィールに基づいた個人支援型の検索を行うとする。

3-4 個人プロフィール部

本部分は「個人プロフィール」から成り立つ。この部分は協調フィルタリングを行うための単語の登録を行い、その単語の評価値なども一緒に保存しておくとする。

しかし、実際検索を行った場合はユーザの嗜好が長年趣味としている事に関連した長期的に使われる単語や瞬間的 (TV で気になった歌手) に連想された短期的に使われる単語があると考えられる。そこで、個人プロフィールには短期キーワード・データベースと長期キーワード・データベースを分けて使おうとする。この2つのDBの分け方についてだが、その日数 (初めて登録された日から) によって、短期キーワードと長期キーワードとして分けるとする。しかし、それだけではユーザにとって賞味期限が切れた単語が出てきてしまうため、単語の頻度を比べる判定を行うとする。これは短期キーワード内で単語の出た頻度と日数を割った数が 0.5 をきった場合、その単語は削除するといったような方法で行う。

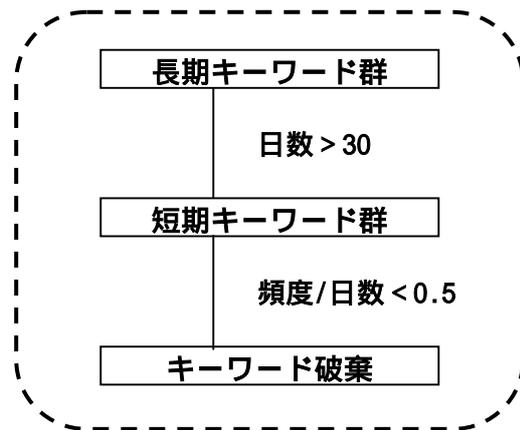


図.2 キーワードDBの分け方

4. 評価

本研究では Web 情報検索の新しい検索方法として、情報フィルタリング技術を用いた Web 情報検索システムを提案した。今までの個人履歴を使った情報検索システムとは違い、個人の嗜好だけにとらわれずに検索結果の網羅性も追求することが出来る。今後は、本システムの実装に向けて進みたいと思う。

5. 考察

従来の情報検索システムは「個人の嗜好」が取り入れられなく、従来の研究では個人の履歴プロフィールだけを使った検索はあったが、それでは「検索の網羅性」が取り入れられない。今回の研究において、情報検索システムにとって大事である「検索の網羅性」と「個人の嗜好」とを両立する事が出来るようになり、より良い情報検索が行えることが出来ると思われる。

6. 参考文献

- [1]: 高橋英史朗「個人履歴情報を用いたWeb情報検索方式の提案」
情報処理学会第65回全国大会, 2004/3
- [2] 大沼・池野, 沖電気工業: 「HTML 文書を対象とした質問応答システムにおける回答抽出方法」,
情報処理学会第 63 回全国大会, 2001/3.
- [3] 仲川こころ・木下敦史: 「対話的に調整可能な文書ランキング WWW検索支援の一手法」
IPSJ, September, Vol143., No9 pp.7-14, (2002)
- [4]: 柘植覚・獅々堀正幹「サポートベクターマシンによる適合性フィードバックを用いた情報検索」
IPSJ, January, Vol144., No9 pp.59-67, (2003)