5T-1

ユーザの探索行動に応じた Web コンテンツ推薦システムの研究

西田　要介†　　東　基衛†
†早稲田大学大学院　理工学研究科　経営システム工学専攻

1. 研究背景

近年，インターネットの急速な普及により，膨大な Web コンテンツが存在し, ユーザは情報収集の際に Web 上の情報を利用する場面が増えてきた．その中から必要な情報を獲得する際，検索エンジンが用いられることが多い．しかし，必ずしも最適な情報が得られるとは限らない．一方，情報検索における行動履歴にはユーザ興味に関する情報が含まれていると考えられ，それに関する研究が活発に行われている[1]．しかし，ユーザの興味は非常に流動的で把握するのが困難である．

本研究では，検索エンジンを利用した検索行動を一つの Web 上の情報検索行動と捉える．その中で変化していくユーザの興味を捉え，必要な情報提供を実現することで, 効率的なブラウジング支援を可能とする Web コンテンツ推薦システムを提案する．

2. 研究アプローチ

効率的な情報収集支援を実現させるためには, ユーザの興味を適切に抽出する必要がある. そこで, ブラウジングにおける閲覧履歴が記録されたアクセスログから得られるデータを利用して, ユーザがどのような情報探索行動をとってきたかを把握する. しかし, アクセスログからの情報だけを基に閲覧プロセス中に興味を惹かれた内容を推測することは不可能である. したがって, Web コンテンツの html ソースであるテキスト文を抽出して, 自然言語処理を行い, 単語の出現頻度を基に Web コンテンツを特徴付けることによって, ユーザのブラウジングにおける興味を把握する.

本研究では，特に検索エンジンの検索結果から興味あるコンテンツをブラウジングする過程で，ユーザの興味が絞り込まれていくことに着目した．その際，本研究では検索結果ページの頻出単語（広域的興味）と，ブラウジング時の Web コンテンツから得られる頻出単語の違いを考慮することで狭域的興味を抽出する．さらに，狭域的興味と関連深い語を単語間の共起頻度に着目し抽出する．それによって，ユーザの絞り込まれる興味変化に対応したブラウジング支援が可能となる．

Web-Contents Recommendation System based on User Search Behabior
Yosuke NISHIDA†, Motoei AZUMA†
†Dept. of IMSE, Graduate School of Sci. & Eng., Waseda Univ.

3. 提案手法

3.1. 提案手法の要件

前述の研究アプローチで顕在化した提案システムの要件を以下にまとめる.

1. ブラウジングにおける閲覧履歴を取得し, 閲覧した Web コンテンツの情報を取得，解析する.
2. 閲覧した Web コンテンツをもとにブラウジングにおけるユーザの興味変化へ対応させる.
3. 適切なキーワードを提示し, WWW 全体からユーザの興味に関連する Web コンテンツを取得するために，検索エージェントを利用する.

3.2. プロキシサーバのアクセスログから得られるデータの利用

本研究では閲覧履歴の取得に際して，プロキシサーバのアクセスログを用いる．プロキシサーバはプラットフォームに非依存，設定が容易，閲覧対象に制限がない，などの利点がある．プロキシサーバから取得できるデータのうち，タイムスタンプ，クライアントアドレス，URL，コンテンツタイプを利用し，ユーザと閲覧した Web コンテンツの情報との関連を明確にする.

3.3. ユーザの興味抽出手法

Web ページは，その多くが何らかの文字情報を含んだテキストである．ブラウジング時におけるユーザプロファイルはテキストから抽出したキーワードと重みからなるベクトル空間モデルで表現することができる.

そこで，本研究で，予備実験を行ったところ，未知の分野を調べたい時，例えば「Java」に関して調べたいユーザはまず検索クエリを「Java」と入力し，結果ページを表示させる．その中から興味を持ったページを選択後，興味に従って，リンクを辿って Web コンテンツを閲覧する．その際、「変数」や「Eclipse」など，少しずつ興味の範囲を狭めながら Web 情報探索行動を行う傾向が見られた．このように未知の分野を調べる際，まず求める情報の核として絶対に外す事のできない語のみを用いて大雑把に検索を行い、その結果を見ながら徐々に入力語を変更・追加して結果を絞り込んでいく，という検索スタイルが好まれていると考えられる.

そこで，本研究で定義した情報検索行動を Google 検索結果ページとそれ以降の Web ページ閲覧ページを分け，頻出語の差異，広域的興味から狭域的興味へ興味の範囲が絞り込まれ，変化していくことに着目した.

ここで，閲覧した Web コンテンツ中に含まれる単語を w_k とする．検索結果ページ中の得られる単語の出現頻度を要素とするユーザの広域的興味 $\vec{S}$ と現在のブラウジングにおいて閲覧した Web コンテンツ中の単語の出現頻度を要素とする狭域的興味 $\vec{T}$ は，以下の (1) 式，(2) 式で与えられる．

$$\vec{S} = (s_1, s_2, s_3, \cdots, s_m) \quad (1)$$

$$\vec{T} = (t_1, t_2, t_3, \cdots, t_n) \quad (2)$$

ここで，t_k（k=1,2,3,・・,n）と同一の要素を有するものを $\vec{S}$ から抽出して作成される $\vec{S'}$ は以下，(3)式で表すことができる．

$$\vec{S'} = (s'_1, s'_2, s'_3, \cdots, s'_n) \quad (3)$$

ブラウジング中のユーザの狭域的興味 $\vec{T'}$ は $\vec{S}$，$\vec{S'}$，$\vec{T}$ を用いて，以下(4)式で表すことができる．

$$\vec{T'} = (t'_1, t'_2, t'_3, \cdots, t'_n) = \left\{ \log \left(\frac{\sum_{l=1}^{m} s_l}{s'_k} \right) \right\} (t_1, t_2, t_3 \cdots t_k \cdots t_n) \quad (4)$$

ここで，ブラウジング中にユーザの一時的興味を示す単語 w_k と共起した単語を W とする．単語 W と狭域的興味群との共起の偏りに狭域的興味語の重み t'_k を乗じたものの総和を単語 W の興味度 D として表す．仮に単語 W と狭域的興味群の間に何らかの意味的なつながりがあれば，単語 W と狭域的興味語との間に共起の偏りが生まれ，D の値は大きくなるはずである[2]．従って，D の値が大きい単語 W をユーザの狭域的興味語と関連のある語の候補とする．

ユーザの狭域的興味語 w_k 単独での生起確率を p_{w_k} とし，単語 W と狭域的興味との共起の総数を Cw，単語 W と単語 w_k の共起頻度を $freq(W, w_k)$ とする．Cw，D は以下の（5）式，（6）式で表すことができる．

$$C_w = \sum_{i=1}^{n} freq(W, w_i) \quad (5)$$

$$D(W) = \sum_{k=1}^{n} t'_k \left(freq(W, w_k) - C_w p_{w_k} \right) \quad (6)$$

3.4. Google Web APIの利用

本研究では，WWW 上の情報源から Web コンテンツを推薦することを可能にするため，Google Web API を用いた検索エージェント機能を実現する．

また， Google Web API を用いることで Google の持つ世界中の豊富で膨大な量の Web コンテンツを対象として情報を取得することができる．

4. プロトタイプの実装

前述の提案手法に基づいてプロトタイプを実装した．以下にプロトタイプの概要を示す．

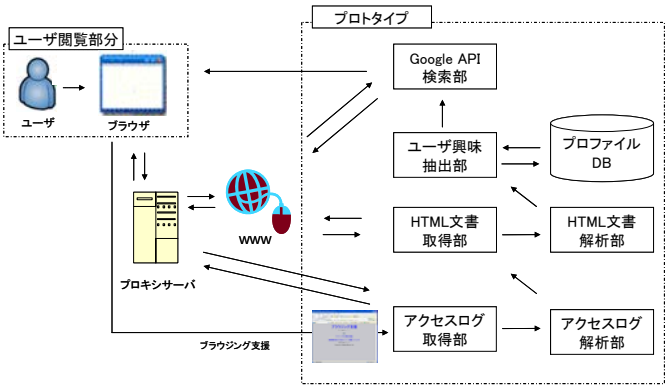

図 1 プロトタイプの概要

プロトタイプは以下のフェーズに従い処理を実行する．

アクセスログ取得部

ユーザがブラウジング中に閲覧したアクセスログをプロキシサーバから取得する．

アクセスログ解析部

得られたアクセスログから必要項目を抽出する．

HTML文書取得部

アクセスログから取得した URL から HTML ソースを読み込む．

HTML文書解析部

HTML ソースを形態素解析して，名詞を抽出し，ストップワードを削除する．

ユーザ興味抽出部

検索結果ページと閲覧履歴を基に，ユーザの広域的興味と狭域的興味を算出し，単語間の共起頻度を元に狭域的興味語と関連する語を抽出する．

Google API検索部

狭域的興味を表す検索クエリとして WWW から情報取得し，出力結果を推薦 Web コンテンツとして提示する．

5. おわりに

本研究では，Web 情報検索行動における閲覧履歴からユーザの狭域的興味とその関連語を抽出し，検索エージェントを利用して Web コンテンツを推薦するシステムを提案した．これにより，Web 情報探索行動で絞り込まれていく興味に対応した効率的な情報収集活動の支援を実現できた．

参考文献

[1]. 松尾豊：”ユーザの個人の閲覧履歴からのキーワード抽出によるブラウジング支援” 人工知能学会論文誌 Vol.18 No.4 E，pp.203-211，2003

[2]. 橋本雅幸:”興味の派生を考慮した Web コンテンツ推薦システムの提案” 情報処理学会全国大会，2005