6C-2

Proxy Logに基づいたコンテンツ自動推薦による知識共有支援システムの提案

丹 英之† 本田 光太郎† 芝崎 亮† 山口 哲† 千葉 大作† 原 誠一郎†

{tanh,hondako,shibasr,yamagust,chibad,haras}@alpha.co.jp

株式会社 アルファシステムズ†

1. はじめに

近年，Google を代表とするインターネット検索エンジンの発達により，個人のみならず組織でも必要とする知識をインターネット上に求めることが多くなった．ところが，インターネット上の知識は無秩序に分散されており，目的とする知識へ回り道をすることなく辿り着けるかは，各人の情報検索能力に左右される．

組織とは，一定の共通目標を達成するために，成員間の役割や機能が分化・統合されている集団のことである．そして，その目的達成のための専門領域を示す概念体系を持っていると考えられる．それ故，この概念体系は，各成員が閲覧した Web ページ集合にも反映されると想定できる．

さて，組織の成員が行う Web 閲覧は，大抵の環境において Proxy を経由して行われる．つまり，インターネットから組織内へ流入する知識の記録である Proxy のログをマイニングすることによって，組織が要求している知識の傾向を把握できると考えられる．そこで，成員の Web 閲覧履歴である Proxy のログを基に，他成員が必要とするであろう知識(コンテンツ)を含む Web ページを，各成員の Web 閲覧履歴から自動的に推薦しあうことで，目的の知識へ短時間で到達でき，且つ，成員が獲得できる知識の均一化を図る方法を検討することにした．

本研究における目標は，組織の成員であるユーザが必要とするコンテンツが含まれるページ集合を推薦しあう仕組みの構築である．このコンテンツの自動推薦に際し，まず，Web 閲覧履歴から推薦するに価するコンテンツの所在を抽出できるかについて事前評価を行った．本稿では，この試みについて述べる．

2. 推薦に価するコンテンツと Web 閲覧履歴

Amazon.com を代表とする EC サイトでは，コンテンツである商品の推薦に協調フィルタリングなどの手法を用いることで商業的にも成功している．この手法は，対象とする顧客と類似する購買行動をとる顧客を抽出することによって，対象とする顧客の嗜好を推測する方法である．そして Web 閲覧においても，対象となるユーザと類似した Web 閲覧行動をとるユーザを抽出することによって，顧客に対する嗜好，言い換えればユーザが要求するであろうコンテンツを指す URL を推測することに他ならない．つまり，ユーザ間で Web 閲覧行動の類似度を評価すれば推薦するに価するコンテンツが分かることになる．

Proxy のログには，クライアントの端末名，時刻，参照先 URL が記録されており，このままでは扱いにくい．そこで各ユーザの閲覧履歴からベクトル空間モデルを構築し，各ユーザ間の関連及び，時間発展について評価することにした．まず，URL 中に含まれるドメイン名が一つのカテゴリーに属するコンテンツを提供すると仮定し，ユーザ a の Web 閲覧履歴の特徴ベクトル$\vec{U}_a$を以下の様に定義する．

$$\vec{U}_a = \left(d_{a1}, d_{a2}, \cdots, d_{aj}\right)$$

ここで，d_{a1}はユーザ a が参照した URL 中のドメインd_1から GET メソッドでファイルを取得した回数である．また j は特徴ベクトルの次元数，すなわち Proxy がアクセスしたユニークなドメインの総数である．次に，ユーザ a, b 間での Web 閲覧行動特性を比較するため，以下のコサイン相関値によって類似度を求める．

$$Simirarity(U_a, U_b) = \frac{\vec{U}_a \cdot \vec{U}_b}{\left|\vec{U}_a\right|\left|\vec{U}_b\right|}$$

得られたユーザ間の類似度を基に R[1]を用い，多次元尺度構成法(MDS)によって各ユーザの類似的な位置関係を把握する．

3. Proxy Log の収集

実験協力者に，ログ収集用 Proxy を経由して休憩時間を含めた業務時間中の Web 閲覧を通常通り行ってもらった．ログ収集の期間は 20 週間で，合計 23 人が参加した．

ユーザが必要とする知識は文字情報から得られるとし，拡張子判断で JPEG, GIF, PNG などの画像，及び，RSS, RDF などサイト更新情報配信ファイル，明らかに広告サイトと判るドメイン，そしてイントラネット内サーバへのアクセスを除去した．各ユーザは Google を利用することが非常に多かった．このため予備実験では，高次元で且つ要素が殆ど零のスパースなベクトル$\vec{U}$の特徴は，Google の成

"Proposal of knowledge sharing system using contents recommendation based on proxy log analysis"

† Alpha Systems Inc.

分に引き摺られ埋もれてしまった．そこで，Google へのアクセスも除去することにした．これらクリーニングによって，期間中の GET メソッドのリクエスト総数 2,324,427 中，有効リクエストは 430,901 となった．また特徴ベクトル$\vec{U}$の次元数である総ドメイン数は 11,744 であった．

各ユーザの特徴ベクトルは，1 期あたり 4 週間とし，5 期に分けた Web 閲覧履歴から求めた．

4. 結果と考察

第 1 期分のログを処理し各ユーザの Web 閲覧履歴の類似的位置関係をプロットしたものを図 1 に示す．各ユーザは三つのグループ A,B,C に分かれる傾向を示した．グループ A には主に“@IT”，“linux.or.jp”，“e-Words”など比較的ドキュメントが揃っているサイトを参照するユーザが集まった．ユーザ特性としては，勤続年数 1 年未満が多く見られ，プロジェクトへ新しく配属され Web で調査しながら業務を進めていることが伺われる．また，A に分類されたユーザ間では，Linux を用いてサーバを用意するなど，プロジェクト内で割り当てられたタスクにも類似性が見られ，関わるプロジェクトが異なっていても，ユーザが要求する知識であるコンテンツの所在を指す URL は，タスク単位で取り扱うことができると考えられる．一方，B,C は勤続年数 1 年以上のユーザが占めており，グループ B,C 間の違いは，情報収集のためよく閲覧しているであろうと思われるニュース系サイトの違いに拠るものであった．勿論，B,C に属するユーザが，A が参照しているサイトをまったく閲覧していないというわけではなく，単に閲覧する度合いが低いだけであり，逆もまた然りである．

5 期に渡って各ユーザの類似度を比較した結果，各ユーザの特徴ベクトルは常に一定ではなく，時間経過と共に揺らぎが見られた．これも，各々のユーザが担当するタスク，そして休憩時間での Web 閲覧からであろう各ユーザの興味の移り変わりに由来すると考えられる．また，ある時期のユーザの特徴ベクトルが，他ユーザの過去の特徴ベクトルに類似する場合があることが多々見られた．これにより，誰かが以前よく閲覧していたサイトを発見し，それを閲覧する機会が増えるということが繰り返されていると考えられる．つまり，誰かが閲覧していたサイトにある知識を他者が必要とする場合があると言える．

5.おわりに

Proxy のログから各ユーザの Web 閲覧履歴を抽出し，URL 中のドメイン名を用いた特徴ベクトルの類似度の比較により，Web 閲覧行動が類似しているユーザが組織内にも存在することを確認した．よって，組織内で各ユーザの Web 閲覧履歴から推

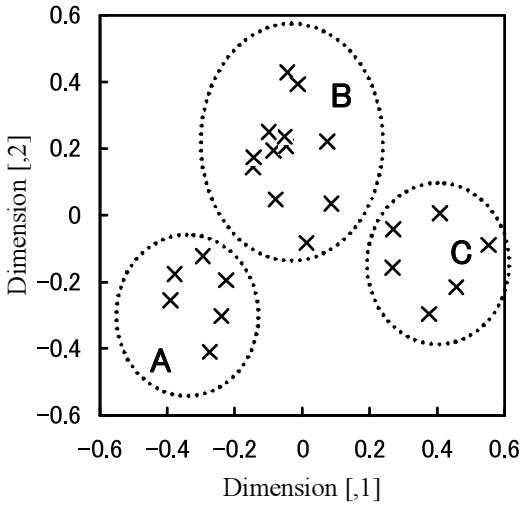

図 1 第 1 期の各ユーザの Web 閲覧履歴特徴

薦に価するコンテンツの所在を抽出できると言える．また，時間経過と共にユーザの特徴ベクトルは揺らぎ，他ユーザが過去によく閲覧していたサイトを参照するようになることが見られた．これは，閲覧履歴をある期間単位で区切ることによって，推薦する知識のカテゴリーを増やすことが可能であることを示している．

今回の評価では，一つのサイトが単一のカテゴリーの知識を提供すると仮定し，URL 中のドメイン名のみを扱った．このままでは，粒度が大きく推薦精度を期待できない．そこで，パスに含まれる文字列からのコンテンツのメタ情報抽出[2]や，検索エンジンに投入されるクエリー文字列,そしてコンテンツの内容も含め，より粒度の小さいデータを扱う Web 内容マイニングへと展開させたい．また Web コンテンツを扱う場合，コンテンツの質が課題となる．これには，自動推薦の精度の問題と同じく，推薦閾値の判断に人手を介すなどの工夫[3]が必要になる．そこで Proxy の利点を活かし，ページの閲覧時間取得や，ユーザの評価をフィードバックする簡単なインターフェースの挿入などにより，コンテンツの質を評価する仕組みを検討したい．そして得られた知見を基に，組織内の知識共有を支援するシステムの設計を行っていく．

参考文献

[1] The R Project for Statistical Computing http://www.r-project.org

[2] 田村剛士 “Web 視聴率データからの Web ユーザコミュニティ発見に向けて”,電子情報通信学会技術研究報告, Vol. 102, No. 710, pp. 1-4, 2003

[3] 根本潤, 遠山元道 “閲覧履歴に基づく情報検索の相互支援”, 電子情報通信学会 第 15 回データ工学ワークショップ, 3-B-02, 2004