3A2-5

Proxyで抽出した組織内ユーザのWeb閲覧特徴の時系列変化

Time series variations of Web browsing feature extraction of the user in an organization by proxy log

丹 英之[*1] 本田 光太郎[*1] 芝崎 亮[*1] 山口 哲[*1]
Hideyuki Tan Kotaro Honda Ryo Shibasaki Satoshi Yamaguchi

千葉 大作[*1] 原 誠一郎[*1]
Daisaku Chiba Seiichiro Hara

[*1] 株式会社 アルファシステムズ
Alpha Systems Inc.

We observed series change at the time of Web browsing action feature, by extraction from the log of Proxy for users in an organization. In the result, the attribute of users is correlated with certain feature of Web browsing action. This report discusses analysis of the role in an organization, and the similarity of the feature between users.

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, 個人的な用途のみならず, 企業などの組織内部から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, インターネット上にある知識は無秩序に分散されており, 十分に整理されているとは言えず, 目的とする知識へ短時間で辿り着けるかは, 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, この共通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, 各成員がインターネットから得た知識集合, つまり各成員が閲覧した 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, その組織のネットワーク環境にも拠るが, 大抵の場合 Proxy を経由して行われることが多い. これにより, インターネットから組織内へ流入する知識は, Proxy のログとして記録されると言える. 勿論, メールやチャットなど Web 閲覧以外の経路からも, 組織の外部から知識が流入する. しかし, Web ページのリンクを辿る行動に連動して, 直ちに関連するコンテンツを入手できるという Web の特性は, コミュニケーションを目的とした上記手段とは異なり, ユーザが能動的に知識を入手しようとする目的を成し遂げるために行動を起こす場合が殆どであると考えられる. つまり, 行動を記録したログにおいて, 知識入手を目的とした行動とそれ以外の行動を別けた場合, 前者を示す行動である割合が高くなる. よって, ログのマイニングによって, 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.

本研究における最終目標は, 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. このコンテンツ推薦に際し, まず, Web 閲覧履歴の分析を行った. 本稿では, この試みについて述べる.

連絡先:丹 英之, 株式会社アルファシステムズ,
川崎市中原区上小田中 6-6-1, tel:044-738-4126
mailto:tanh@alpha.co.jp

2. コンテンツの推薦とWeb閲覧の特徴抽出

コンテンツの推薦には, 協調フィルタリングなど様々な方法があり, これらを応用した Amazon.com など大手 EC サイトでは商業的にも成功している. EC サイトでは, 対象とする顧客と類似する購買行動を取った顧客を抽出することで, 対象とする顧客の嗜好を推測し, 商品を推薦している. 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ることで, 必要とするコンテンツが含まれるページを推薦できることになる.

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 閲覧行動の特徴ベクトル $\vec{U}_a$ 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, 次式のコサイン相関値によって類似度を求める.

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

3. Proxy Log の収集と分析

Web 閲覧履歴の取得には, ログ収集用 Proxy を用意し, 所属部署内において実験協力者を募り, 休憩時間も含めた業務時間中の Web 閲覧を通常通りに行ってもらった. 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名，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．そのため予備実験では，高次元で且つ要素が殆どゼロのスパースな特徴ベクトル$\vec{U}$の特徴は，Google の成分に引き摺られ埋もれてしまった．そこで Google へのアクセスを除外することにした．これらの処理により，期間中の GET メソッドのリクエスト総数 2,234,427 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 期当たり 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 閲覧行動の類似的位置関係を多次元尺度構成法で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，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．ユーザへのヒアリングでは，グループ A に分類されたユーザ間では，プロジェクト内での担当タスクにも類似性がみられた．これよりプロジェクトが異なっていても，ユーザが必要としている知識はタスク単位で扱うことができると考えられる．一方，グループ B,C は勤続年数 1 年以上のユーザが占めていた．グループ B,C 間の違いは，情報収集のために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，グループ B,C のユーザが，グループ A のユーザが参照したサイトを全く閲覧していないというわけではなく，単に閲覧の度合いが異なるだけである．

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．ユーザ 8 人，5 期分の組み合わせについて Web 閲覧行動の類似度を示したものが図 2 である．例えば，ユーザ A の第 4 期の閲覧行動は，ユーザ D の全期に渡って類似しており，その時期に必要とした知識が類似した，更には，全期に渡ってユーザ D が担当していたタスクを第 4 期にユーザ A が担当した可能性もある．これら期を跨いだユーザ間にも閲覧行動に類似性が見られることから，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 閲覧特徴を抽出し，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．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時期があることが見られた．よって，組織内で各ユーザの Web 閲覧履歴から，類似した閲覧行動を取ったユーザを抽出でき，そのユーザの閲覧履歴からコンテンツを推薦できると言える．また，時間経過と共に閲覧特徴は揺らぐことから，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．

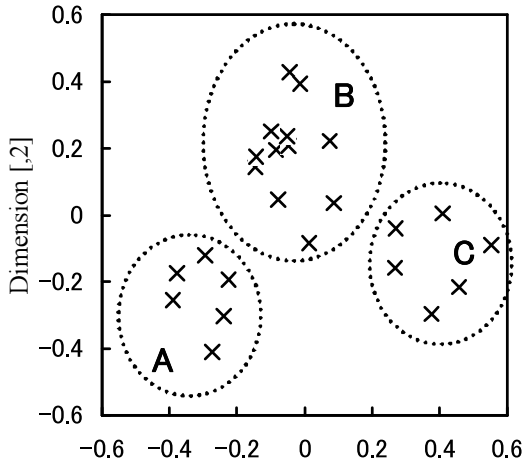

図 1　第 1 期の各ユーザの Web 閲覧行動特徴

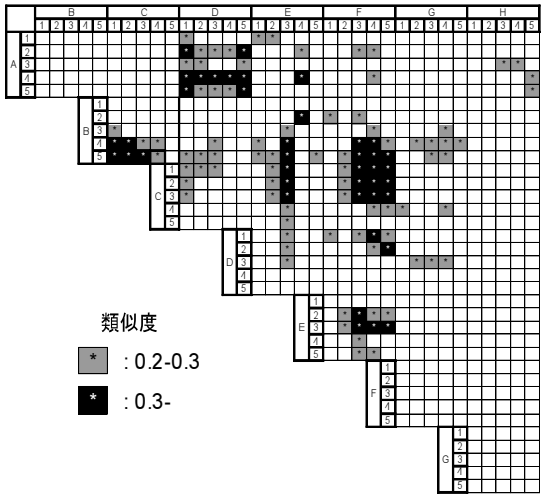

図 2　期を跨いだユーザ間の Web 閲覧行動の類似度

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，粒度が大きくコンテンツ推薦の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[田村 2003]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[根本 2004]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

[田村 2003] 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.

[根本 2004] 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.