動的ページへ適用可能なページ推薦技術の検討

渡辺 拓也† 太田 学‡ 片山 薫‡ 石川 博‡
東京都立大学工学部電子・情報工学科† 東京都立大学大学院工学研究科‡

１．はじめに

インターネットサイトの肥大化に伴い、求めるデータに到達するまで多くの労力が必要になっている。この問題への解決策として、ページ推薦技術が開発されており、その例として SiteHelper[3]等がある。しかし、SiteHelper は被推薦者にフィードバックを求める事で新たな作業を要求してしまうので、本研究ではこのアプローチは取らない。

本研究では単一サイトのログデータのみから推薦の元になるデータを作り、それを用いてそのサイトを訪れた人に推薦を行う方法を考えた。

２．頻出パスによる推薦

２．１．推薦の方法

サイトユーザーがサイトを訪れてから去るまでの１連のＵＲＬリクエストの並びを１トランザクションと呼ぶ。いくつかのＵＲＬの並びをパスと呼び、（パスを構成するＵＲＬの個数－１）をパスの長さとする。あるパスを含むトランザクションの個数をそのパスの support と呼び、support(A)と書く。あるパスＡを含むトランザクションのうちｃ％がパスＢを含む時、ｃをパスＢのパスＡに対する confidence と呼び、confidence(A → B)と書く。

サイトのログデータから頻出パスを求め、「被推薦者の様にサイト内のページを見てきた人が次によく見ているページ」を推薦する。例えばａ→ｂ→ｃ→ｄ→ｅ→ｆという頻出パスがあり、被推薦者がｂ→ｃ→ｄとブラウジングした時､ｅとｆを推薦する。本研究では推薦されるページが多くなりすぎて被推薦者が混乱する事を避けるため、頻出パスと被推薦者のブラウジングしたパスとの重複部分の長さが２以上のパスのみ推薦した。以下、この重複部分の長さを重複度と呼ぶ。

頻出パスの生成アルゴリズムは apriori アルゴリズム[5]を参考にした。集合演算を行うため、トランザクションにはそれぞれＩＤを付加する。

まず、ログファイルから全ての長さ１のパスについて support を求める。support がある定数（minsup）を越えている物のみを残す。次に、残った長さ１のパスを組み合わせて生成しうる全ての長さ２のパスについて support を求める。例えばある２つのパスがそれぞれａ→ｂ、ｂ→ｃというパスであるなら、ａ→ｂ→ｃというパスを生成しうる。ここで、ａ→ｂ、ａ→ｂ→ｃというパスをそれぞれX、Yとし、support(Y)、confidence(X → Y)を求める。そしてsupport(Y)、confidence(X → Y)が共にある定数 minsup,minconf を越えているパスYのみを残す。長さ２の２つのパスａ→ｂ→ｃ、ｂ→ｃ→ｄからはａ→ｂ→ｃ→ｄという長さ３のパスを生成しうる。この様に計算するパスの長さを１ずつ増やしていき、計算するパスがなくなるまで操作を繰り返す。この計算の課程で、長さ２以上のパスの生成は集合演算を行うだけで済むので、ログファイルへのアクセスは長さ１のパス生成時に１度行うだけである。

２．２．推薦強度の算出方法

被推薦者にページを推薦する際に､推薦の強度(以下推薦値と呼ぶ｡）を一緒に提示した方が親切である。本研究では頻出パスの support そのものを推薦値としたが、頻出パスを生成したトランザクションの日付、曜日、時間帯などによって重み付けをする方法も考えられる。また、support を推薦値計算に使用するので、１つの頻出パスに含まれる長さ３以上のパスそれぞれの support を頻出パス生成時に保存する。

また、頻出パスとの重複度が大きくなるにつれ、頻出パスは被推薦者の特性をより強く表していると考えられるので、推薦値も大きくなるべきである。重複度と推薦値を別次元の物と考え、両方を被推薦者に提示する方法も考えられるが、わかりやすさを考え重複度を推薦値計算に利用し、推薦値のみを提示する。同じ推薦ページでも、重複度や頻出パスが異なるとそれぞれについて推薦値が計算されるが、この内最も値の大きい物をその推薦ページの推薦値とする。

さらに、頻出パスを含むトランザクション数にトレンド[2]がある場合、推薦値を強調する方法が考えられる。単純に正負のトレンドを示すアイコンを提示するだけでもよいが、推薦値計算にトレンドを使用する方法を考える。

トレンドを計算する単位として、週単位と日単位を考えた｡週単位は曜日による変動を除去できるが、長期のトレンド検出しかできない。日単位は短期のトレンド検出ができるが､曜日による影響を受ける。曜日による影響を除去するため、土（日）曜日のトランザクション数に（サイト全体の平日の平均トランザクション数／サイト全体の土（日）曜日のトランザクション数）をかければ曜日による影響を排除できると考えられるが、誤差が生じてしまう。

トレンドのあるパスは過去何単位かの support 変移から support を数式で近似し、１単位未来の予測 support を support とする方法を考えた。トレンドの検出にはケンドールによるトレンドの検定の公式[2]を使用した。

２．３．動的ページへの応用

論理値、整数、文字列を環境変数として使用する動的ページは、同じ様な内容のページを表示していてもＵＲＬが違うので全く違う物として扱われてしまう。よって、環境変数から動的ページを含むパスをグルーピングし､グループを同じパスとして扱う。しかし、グルーピングされたページは特定のページを示してはいないので、頻出パスに含まれていても推薦対象にはならない。被推薦者が動的ページを通った時に推薦ページを導出するのに使用する。

グルーピングの方法として、apriori アルゴリズムを参考にした。動的ページが扱う変数１つ１つについてグルーピングを行い、そのグループを含むトランザクション数が minsup を越えているグループのみ残す。次にグループが扱う変数の数を１増やし、考えられる組み合わせについてグルーピングを行い、グループの含むトランザクション数が minsup

---

A Page Recommendation Techniques applicable to dynamic pages.

Takuya Watanabe †, Manabu Ohta ‡, Kaoru Katayama ‡, Hiroshi Ishikawa ‡
† Electronics and Information Engineering, Faculty of Engineering, Tokyo Metropolitan University.
‡ Graduate School of Engineering, Tokyo Metropolitan University.

を越えているグループのみ残す。この操作を新たな組み合わせが作れなくなるまで繰り返す。

ここで、動的ページが範囲の狭いグループに属する時､より強く特性を表していると考えられるので、support が minsup を越えた場合は（全体に対するグループのしめる領域の割合の逆数× support）を support とする。ただし、このままでは動的ページの support が大きくなりすぎてしまうので、動的ページと静的ページ間で整合を取る事が求められる。

３．頻出パスを使用しない推薦方法

頻出パスを使わない推薦方法として、ある２つのページの共参照されやすさを求め、被推薦者の見てきたページと共参照されやすいページを推薦する方法を考えた。まず、ページリクエスト毎にページを「よく見た」又は「それ以外」というカテゴリーデータを持たせる。閲覧時間はログファイルに記録されたリクエストの間隔より求める。ここで実際の閲覧時間ではなくカテゴリーデータを使用するのは、最後にアクセスしたページの閲覧時間が算出しにくい事、なんらかの理由で極端に長時間閲覧されたデータの影響を小さくするためである。閲覧時間をカテゴリーデータに変換する際、ページ毎に閲覧時間の上半分を「よく見た」とした。そして、カテゴリーデータのまとまりから多変量解析の数量化Ⅲ類[1]を使用して各ページに何次元かの値を割り当て、被推薦者の見てきたページと「近い」ページを推薦する方法を考えた。

４．実験

本実験では東京都立川市にある映画館のＨＰの２００１年２月～１０月（９月はのぞく）のログデータを用いた。

４．１．頻出パスの生成

アクセスログから画像ファイルへのアクセスやアクセスに失敗した物を除くと 1694100 アクセスが残り、そこから 871714 トランザクションが生成された。さらに、その内長さ３以上の物は 155790 であり､そこから長さ３以上の頻出パスが 6 生成された。

４．２．トレンドを用いた推薦

２００１年の６月のログファイルから週単位で各頻出パスについてトレンドを測定し、３種類の数式で１週間未来のデータを予測した。全ての頻出パスについて正のトレンドが見られ、予測の結果は表１のようになった。

表１．トレンドを用いた予測値

1	2	3	4	5a	5b	5c
1	5	20	49	59	90	219
0	3	5	6	9	6	
0	3	5	6	9	6	
1	5	8	9	13	9	25
1	4	7	9	12	11	24
2	4	10	11	15	14	25

各行はそれぞれ頻出パスを表す。１～４は１週目から４週目のログ数で、5a ～ 5c はそれぞれ A+BT、$A'+B'T+C'T^2$、$A''e^{B''T}$ で５週目のデータを予測した物である（A,B,A',B',C',A",B"は定数、T は時間)。

5a はほぼ等しい割合で値を増加させ、5b は収束しつつあるトレンドについてはあまり変化を見せず、5c はトレンドの強さをより大きく反映する。

４．３．数量化Ⅲ類を用いた推薦

データ量を抑えるために１０月のログデータのみを用いた。36 種類のページから成る､「よく見た」ページを１つ以上含む 53756 トランザクションから数量化Ⅲ類を行った。最も有意と思われる３つの値の組を３次元値としてとらえ、ユークリッド距離を計算し、大小３番目まで、ページのＩＤの組とユークリッド距離を表２に示す。この距離は２つのページが同時に参照されやすい程近くなる物で、以下ページ間距離と呼ぶ。ページとページに何の関連もない時２つのページａとｂを共に参照するトランザクションの数をＺとすると、

Ｚ＝ａを含むトランザクション数×ｂを含むトランザクション数／全トランザクション数

と、なる。それぞれのページを含むトランザクション数とＺを表３に示す。表２、表３より、今回求めたページ間距離は、「表３の②」、又は、「Ｚに対する表３の②の相対値（②／Ｚ)」が大きくなるとページ間距離が小さくなるという事に矛盾しない事がわかる。

この距離を用い、被推薦者の過去２，３個の閲覧履歴より、最短距離法を用いて推薦を行う。

表２．ページ間距離が大きい又は小さいページの組

	ページの組	ページ間距離
最大値	(14,202)	7.17×10^{-2}
２番目に大きい	(33,202)	6.91×10^{-2}
３番目に大きい	(30,202)	6.87×10^{-2}
最小値	(68,69)	7.48×10^{-5}
２番目に小さい	(197,261)	1.52×10^{-4}
３番目に小さい	(70,72)	1.65×10^{-4}

表３．特定のページを含むトランザクション数

①	②	Z
202	337	
14	2307	
33	8994	
30	11538	
14,202	13	14.5
33,202	38	56.4
30,202	45	72.3

①	②	Z
68	2518	
69	414	
197	9837	
261	6716	
70	527	
72	444	
68,69	317	19.4
197,261	2925	1229.0
70,72	136	4.4

①はページＩＤ、②は①のＩＤのページを含むトランザクション数

５．おわりに

本研究ではサイトのログデータから頻出パスまたはページ間距離を求め、ページ推薦を行う方法を考察した。

今後、動的ページを扱うプログラムを実装し、実際にページ推薦を行い、頻出パスと数量化Ⅲ類を使用した推薦の評価をしたい。また、同様の実験で重複度とトレンドの利用効果を評価したい。

謝辞

本研究の一部は文部科学省科学研究費特定領域研究(2)[情報学:A02](課題番号:14019075)による。

参考文献

[1]有馬哲　石村貞夫、多変量解析のはなし（東京図書,1987 年）

[2]石村貞夫、グラフ統計のはなし（東京図書,1995 年）

[3]Daniel Siaw Weng Ngu and Xindong Wu, SiteHelper:A Localized Agent that Helps Incremental Exploration of the World Wide Web.In *6th International World Wide Web Conference*, Santa Clara, CA,1997.

[4]Bamshad Mobasher et al,Creating Adaptive Web Sites Through Usage-Based Clustering of URLs.In *IEEE Knowledge and Data Engineering Workshop (KDEX'99)*, 1999.

[5]Rakesh Agrawal and Ramakrishnan Srikant,Fast Algorithms for Mining Association Rules.In *Proc. of the 20th VLDB Conference*, pages 487-499,Santiago, Chile, 1994.