

ベイジアンネットによる Web ブックマーク選択モデルを用いた ブックマーク推薦法

相良 光志[†] 北越 大輔[†] 鈴木 雅人[†]

[†]東京工業高等専門学校

1. はじめに

Web ブックマークは繰り返し閲覧する Web ページを登録する機能であり、Web ブラウジング（以降ブラウジング）の効率を上げるために活用される。一方で、その登録件数の増大に伴い、膨大な選択肢の中から必要なブックマークを発見することが困難となり、かえってブラウジングの効率が低下する可能性がある。

本稿では、ブックマークを効果的に活用した快適なブラウジングの実現のため、確率モデルの一つであるベイジアンネット（Bayesian Network: BN）を用いたブックマーク推薦法を提案する。BN は他の確率モデルと比較して不確実性・動的性を含む環境に対して有効であるため、ブックマーク推薦に適用することで、ユーザの状況（例：仕事、休憩等）を捉えるだけでなく、仕事によく利用するページを閲覧した後、休憩のため、娯楽に関するページを閲覧する、といったような利用状況の変化に対応した推薦が可能であると考えられる[1]。提案手法では、ブックマークがいくつかのフォルダ（例：音楽、ニュース等）に分類されていることを前提とし、ユーザの過去の Web ページ選択履歴を用いて構築した BN（ブックマーク選択モデル）より確率推論を行なう。これによって、ユーザの数回の Web ページ選択履歴から次に選択される確率の高いフォルダを推薦する。本稿では数値実験によって、提案手法の基本的特性を評価する。

2. ブックマーク選択モデル

BN は、確率変数（ノード）間の依存関係を非循環性有向グラフで表した確率モデルの一つである[2]。2つのノード間の依存関係が存在する場合、当該ノード間には有向リンク（矢印）が描かれる。BN 中のノード間の関係性は条件付確率で表現され、結合構造を用いた確率推論によって、不確実性を含む事象の予測が可能である。本稿では、ユーザの Web ページ選択履歴から

A Study on Web Bookmark Recommendation Method using
Bookmark-Selection Model

[†]Koji Sagara, Daisuke Kitakoshi, Masato Suzuki, Tokyo
National College of Technology

表1 ブックマーク選択モデルのノード

ノード	ノードの意味	取りうる値
$F_i(n)$	フォルダ i を n 回前に選択	0,1
T	選択された時間帯	0,1,2,3
D	フォルダが選択された日	0,1

$i = \{x | 1 \leq x \leq I, x \in \mathbf{N}\}$ (I : フォルダ数)

$n = \{y | -1 \leq y, y \in \mathbf{Z}\}$

BN（ブックマーク選択モデル）を構築する。ブックマーク選択モデルはブックマークの選択される順番・日時の関連性を表現し、表1に示すノードによって構成される。表中の $F_i(n)$ において、 $n=0$ は現在、 $n=-1$ は次回を意味する。 $F_i(n)$ の値は、当該フォルダが選択された (1) か否か (0) に対応し、 D の値はそのフォルダが休日に選択された (1) か否か (0) に対応する。また、 T の値 (t) は $6t$ 時から $(6t+6)$ 時の間に選択されたことを表している。BN の構造は、ブックマーク推薦の各対象ユーザにおいて、ブラウザ起動中にブックマークを選択した順序、フォルダ名、ブックマーク名、日時に関するデータ（Web ページ選択履歴）を用いて決定（学習）される。構造学習には対数尤度を評価関数とした Greedy Search を用いる。

3. ブックマーク推薦法

提案手法におけるブックマーク推薦は、ユーザの数回分の Web ページ選択履歴から、フォルダ選択の有無に関する同時確率変数 $F_i(m)$ ($0 \leq m, m \in \mathbf{Z}$) に対する値 ($f_i(m)$)、および D, T に対する値 (d, t) を定め、BN の結合構造を用いた確率推論法の一つである Loopy-BP を用いて実施される。具体的には、式(1)により求められるフォルダ F を推薦対象として出力する。

$$F = \arg \max_{F_i(-1)} P(F_i(-1) = 1 | D = d, T = t, F_i(m) = f_i(m)) \quad (1)$$

また、定めた値の組み合わせが BN 構築に用いた Web ページ選択履歴に存在せず、推論が不可能となった場合、ランダムに推薦対象を決定する。

4. 実験

東京高専情報工学科 5 年の学生 5 人を対象に実験を行なった。被験者自身が利用しやすいよう、各自の主観でブックマークをいくつかのフォルダに分類してもらった上で 1 ヶ月間ブラウジングを行なうよう依頼し、Web ページ選択履歴を被験者ごとに収集した。

各被験者から収集した Web ページ選択履歴の 1 割をランダムに抜き出してテストデータとし、残りを BN 構築用データとする。データ数が 100 以下の場合には 10-fold 交差検証法を行なう。構築した BN に対し、テストデータを用いて以下の手順で提案手法の特性評価を行なう。

- (1) テストデータから $F_i(m)$, D , T の値を決定
- (2) 手順(1)で定めた値より式(1)を計算
- (3) 手順(2)の結果から得られる F_i が、当該テストデータでも選択されている場合「一致」と判断

各被験者における一致率を推薦精度とする。

5. 実験結果および考察

はじめに、各被験者から収集した履歴データの特徴、および実験期間中に被験者が頻繁に選択した上位 2 フォルダの選択率を表 2, 3 に示す。表 3 より、被験者 A は他の被験者と比較してフォルダ選択の偏りが少ないことが確認できる。また、履歴データの内容を分析すると、一般的に利用状況の変化を捉えづらいデータとなっていることが確認できた。続いて、被験者ごとのブックマークフォルダ推薦精度を表 4 に示す。ここで、各被験者におけるフォルダ選択率と推薦精度を比較すると、被験者 A における推薦精度は最大選択率を大きく上回っており、他の被験者における推薦精度は選択率最大の値とほぼ等しい値となっていることが分かる。各被験者におけるフォルダごとの推薦精度を算出すると、被験者 A の履歴データでは、各フォルダの推薦精度は実際の選択率と同程度かそれ以上の精度

表 2 各被験者の履歴データ

	A	B	C	D	E	平均
データ数	61	113	777	110	1571	526
フォルダ数	7	4	9	8	8	7.2

表 3 各履歴データの上位 2 つのフォルダ選択率

	A	B	C	D	E	平均
第1位	37.7%	80.7%	61.5%	82.7%	53.2%	63.2%
第2位	31.1%	14.0%	16.9%	5.5%	14.6%	16.4%

表 4 各被験者における推薦精度

A	B	C	D	E	平均
42.9%	82.5%	61.5%	83.3%	52.5%	64.6%

を示している。一方、A 以外の履歴データでは、選択率最大のフォルダの推薦精度は約 100% となり、反対に他のフォルダに関する推薦精度は約 0% となっている。これらの実験結果より、被験者 A は、他の被験者と比較してフォルダ選択の偏りが小さく、そのような傾向が効果的な推薦の実現に貢献した可能性がある。

利用状況の変化を捉えづらいデータとなった原因の一つとして、データ収集の事前準備として被験者に行ってもらった「ブックマークのフォルダ分類」において、被験者が実験設定を十分に理解しないまま、利用状況に応じた分類とは異なる分類を実施したことが考えられる。実際に、本来は利用状況が同一（例：娯楽）と考えられるような複数のブックマーク（例：YouTube と YouTube にアップロードされた動画）を別々のフォルダに分類してしまうことで選択率に偏りが生じ、一方のフォルダの推薦精度が著しく低下するという状況が観測された。加えて、普段ブックマークの選択にフォルダを利用していない者にとって、フォルダを利用したブックマーク選択は困難であった可能性が高い。以上の理由から、利用状況の変化を捉えづらいデータとなったと考えられる。従って今後は、提案する手法による推薦の実現に適したブックマークの分類法を事前に提示しつつ、ユーザの利用状況に応じたブックマークの分割・併合の方法などについても検討する必要がある。

6. まとめ

本稿では、快適なブラウジングを実現するため、BN を用いたブックマーク推薦法を提案した。また、数値実験によって提案手法の基本的特性を評価した。被験者が実験設定を十分に理解していなかったため、データのほとんどに大きな偏りが存在した。しかし、選択率の偏りが少ないデータでは効果的な推薦の実現に貢献した可能性があることを確認した。

今後はデータ収集方法、ブックマークの分割・併合の処理について検討する必要がある。

参考文献

- [1] 北越大輔, 塩谷浩之, 中野良平 “利用状況の動的変化に対応するブックマーク情報提示システムの提案”, 第 8 回 AI 若手の集い, 2007
- [2] 本村陽一, 佐藤泰介 “ベイジアンネットワーク：不確定性のモデリング技術”, 人工知能学会誌, pp.575-582, 2010