3H7-3

ベイジアンネットワークを用いたWebレコメンデーションシステムの開発

Development of Web Recommendation System Using Bayesian Network

山崎 敬広[*1] Takahiro Yamazaki　ソンムァン ポクポン[*1] Pokpong Songmuang　石山 洸[*2] Ko Ishiyama　高田 健一郎[*2] Kenichiro Takada　植野 真臣[*1] Maomi Ueno

[*1]電気通信大学 大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications

[*2]株式会社リクルート インターネットマーケティング局
RECRUIT Co., LTD. Internet Marketing Office

This paper introduces an implementation for Web recommendation system using a bayesian network. Bayesian network was used since its prediction is efficient although its calculation amount is heavy. MWST algorithm was used for our system due to the amount of data handled. MWST algorithm gives good prediction, with a calculation amount of $O(n^2)$. Evaluation of the system was applied using the actual data showing its efficiency.

1. はじめに

近年，様々な情報から個々のユーザの嗜好を推測し，コンテンツを推薦する技術の研究が盛んに行われている．筆者らは，大規模 Web サイトにおいてサイト内ページを推薦するシステムを開発中である．推薦の方法として，サイト内の各ページの閲覧確率と依存関係をモデル化し，ユーザごとに推論を行い閲覧確率が高いページを推薦する方式を考えている．本システムでは，多数の確率変数間の複雑な依存関係を柔軟にモデル化できるベイジアンネットワークを用いて，モデル化を行う．

ベイジアンネットワーク [1] とは，予測対象の各変数をノード，変数間の確率依存関係を有向アークとして確率ネットワークを構築したもので，確率構造を表す DAG(Directed Acyclic Graph) と条件付確率パラメータ集合で表現される．ベイジアンネットワークでは，説明変数，目的変数の区別無く，観測された事実を所与として，他の変数の生起確率を，確率理論に裏付けられた計算結果として求めることができる．ベイジアンネットワークの研究は大きく分けて 2 種類あり，１つはデータからのベイジアンネットワークの構造学習，もう１つはベイジアンネットワーク上での確率推論である．ベイジアンネットワークの構造学習の研究では，相互情報量を用いて木を構築する MWST アルゴリズム [2]，ベイズ的アプローチによる予測分布から構造学習を行う K2 アルゴリズム [3]，様々なアプローチから導かれた情報量基準を用いて構造学習を行う手法などが提案されている [4][5]．近年まで，ネットワークの同型性や一致性を持つ手法が，十分なデータ数があれば真のモデルを近似できるとして，盛んに研究が行われてきた．しかしながら，最近の研究で，一致性を持つことが，実際の問題に対して予測効率を最大化するわけではないことが，わかってきている [6]．逆に，古典的な手法である MWST アルゴリズムがデータマイニングの大会などで成果を挙げており，その予測効率の良さに注目が集まっている [7]．また，一般的にベイジアンネットワークの構造学習は NP 完全問題であるが，MWST アルゴリズムの場合，計算量が高々$O(n^2)$ となっており，高速に動作することが出来る．また，確率推論の研究では，高速にネットワーク上の全ノードの周辺事後確率を求める手法として Pearl's Message Passing アルゴリズム [8] が提案されている．

本システムでは，サイト内各ページの閲覧回数を変数として，各ページの閲覧確率の関係をデータから学習し，ベイジアンネットワークを構築しようと考えている．ここで重要であるのが，大規模 Web サイトにおける推薦システムとして，(1) 大規模なデータに対して高速に動作し，かつ，(2) 実データに対する予測効率が良いことが求められている点である．これらの点を満たすことから，本システムでは MWST アルゴリズムを用いたベイジアンネットワーク構築を行うこととし，システムに実装した．さらに，構築されたネットワークに対し，大量のデータを証拠とした全ノードに対する確率推論を行う必要があるので，高速化のために，Pearl's Message Passing アルゴリズムを実装した．また，実際に大規模ネットワークを構築し，実データを用いたシステムの予測精度評価を行ったので，これを報告する．

連絡先: 山崎敬広 電気通信大学 大学院情報システム学研究科
〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: yamaza-t@ai.is.uec.ac.jp
ソンムァンポクポン 電気通信大学 大学院情報システム学研究科 〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: pokpong@ai.is.uec.ac.jp

2. ベイジアンネットワークのアルゴリズム

本システムに実装したベイジアンネットワークのアルゴリズムについて説明する．

2.1 MWST アルゴリズム

Chow & Liu による，各変数間の相互情報量を基準として木を構築する手法である．以下のようなステップからなる．

1. 与えられたデータより，$N(N-1)/2$ 個の枝について，すべての枝の相互情報量 $I(x_i; x_j)$ を求める．
2. もっとも大きな相互情報量を示す枝を取り出し，木を構成する枝とする．
3. 次に大きな相互情報量を示す枝を木に加えるが，ループが出来るならその枝を捨てる．
4. ステップ 3 を N-1 個の枝が木に加わるまで繰り返す．
5. 最後に根ノードを決定し，根から葉へ向かうように枝に方向を付ける．

このアルゴリズムの利点として，以下が挙げられる．

- 二次統計量までしか用いないため，データからの演算が容易で信頼できる．

- 計算量がノード数 n に対して高々$O(n^2)$ である．

このため，ノード数が多いような大規模データに対しても有効に使えると考えられる．また，統計的には一致性がまったくないものの，予測効率が良いと報告されている．

2.2 Pearl's Message Passing アルゴリズム

Pearl による，高速に確率推論を行うための手法である．一般的にベイジアンネットワークでは，ネットワークを構築した後や，ノードにデータを与えた場合，あるノードの事後確率を求めるのに，周辺化を行う必要がある．この周辺化はネットワーク構造を利用して効率的に計算することが出来るが，これを一般化したのが Pearl's Message Passing アルゴリズムである．以下のようなステップからなる．

1. 証拠データを与えられたノードから，その周辺ノードへ向けてメッセージの送信を行う．
2. メッセージを受信したノードは，受信したメッセージを用いて，自分の周辺事後確率を更新する．
3. 周辺事後確率を更新したノードは，メッセージの送信元以外の自分の周辺ノードにメッセージを送信する．
4. 2,3 を繰り返し，全てのノードの周辺事後確率を更新する．

このアルゴリズムを用いると，高速かつ厳密に，全ノードの周辺事後確率を求めることが出来る．しかしながら，ループが存在するネットワークでは，メッセージがループしてしまうため，用いることができない．だが，本システムでネットワーク構造学習に用いる MWST アルゴリズムでは，ループ構造は出来ないため，問題なく用いることが出来る．

3. Web 推薦システムの概要

本節では，開発した Web 推薦システムの概要を説明する．

3.1 システム対象

システムの対象は，膨大な数のユーザを抱える，サイト内全ページ数 10 万以上の大規模 Web サイトである．このサイトのトップページには，様々なジャンルの Web ページへのリンクが張られている．これらの様々なページの中から，ユーザの好みそうなページを推薦し，効率的にユーザを誘導することが本システムの目的である．推薦方法としては，トップページの様々なページへのリンクを並び替え，推薦するページへのリンクを上部に置くような方法を取っている．

3.2 システム概要

Web 推薦システムの概要を図 1 に示す．システムは，ログ取得部，パラメータ演算部，コンテンツ表出部の 3 つの部分からなる．ログ取得部では，ユーザの閲覧履歴を取得し，データベースに格納する．パラメータ演算部では，データベースに格納されたユーザ閲覧履歴を用いてベイジアンネットワークの構築と，各ユーザごとに各ページの閲覧確率の推論を行い，得られた結果をそのユーザの個人パラメータとしてコンテンツ表出部に送る．コンテンツ表出部では，個人パラメータに応じて推薦する Web ページを選択し，推薦するページを強調するようにトップページを構成し，ユーザに表示する．

3.3 ベイジアンネットワークの構築と個人パラメータ導出

パラメータ演算部でのベイジアンネットワークの構築手順と，個人パラメータ導出について説明する．パラメータ演算部は，(1) 全ユーザのページ閲覧履歴からベイジアンネットワークの構築，(2) あるユーザのページ閲覧履歴から各ページの閲覧確率を推論，の 2 つの部分からなる．(1) では，各ユーザが各ページを閲覧したかどうかを全ユーザについて集計し，全ユーザにおける各ページの閲覧確率を求める．この閲覧確率を用いて各ページ間の相互情報量を計算し，MWST アルゴリズムを用いて各ページをノードとしたベイジアンネットワークを構築する．ここで，根ノードとなるページは，最も多くのユーザが閲覧したページとしている．(2) では，あるユーザが閲覧したページをデータとしてベイジアンネットワークに与えて推論を行い，そのユーザにおける各ページの閲覧確率を求める．推論アルゴリズムには Pearl's Message Passing アルゴリズムを用いている．最後に，得られた閲覧確率から閲覧回数の期待値を求め，期待値が大きいページから順に並べていく．これがユーザの個人パラメータとなり，期待値が大きいページをそのユーザに推薦するよう，トップページを構成する．

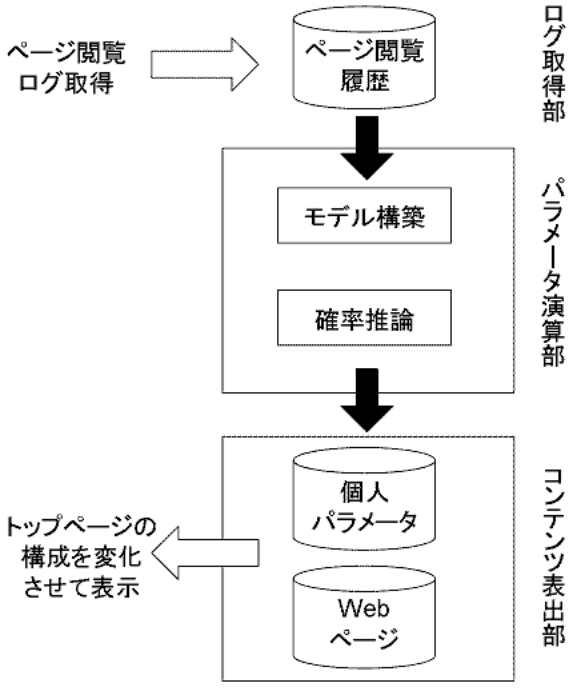

図 1: Web 推薦システムの概要

現状のシステムでは，Web サイト内の一部の Web ページ (5000 程度) について，ベイジアンネットワークを構築し，ページ推薦を行うことができる．サイト内の全 Web ページ (10 万以上) に対するベイジアンネットワークの構築は，計算量が膨大になり実行できないため，何らかの工夫が必要である．現状では，サイト内の各ページを数十個のカテゴリに分け，各カテゴリを変数としてベイジアンネットワークを構築し，ユーザに対してカテゴリを推薦することを行っている．

4. システム評価実験

実際に，ユーザのページ閲覧履歴からベイジアンネットワークを構築し，各ユーザごとに推論を行い，得られた推薦結果について評価を行った．

4.1 実験条件

Web サイト内のあるカテゴリに属す 1475 個の Web ページについて，ベイジアンネットワークを構築し，実験を行った．実データから，15612 人のユーザの閲覧履歴をサンプリングデータとして用いた．15612 人のユーザデータのうち，5612 人のユーザデータを学習用データとして構造学習を行い，残りの 10000 人のうち，ページを 10 種類閲覧しているユーザ 102 人をテストデータとして，推論を行った．各ページの閲覧回数は，0,1,2,3,4 回の 5 つに離散的に設定し，閲覧回数 4 回以上の場合は 4 回として扱った．

4.2 評価手法

ユーザのページ閲覧履歴のいくつかを用いて推論を行い，その結果求められる推薦結果と，推論に用いた以外のユーザが閲覧したページとが，どれだけ合致するかに着目して評価を行った．具体的には，推論で得られた各ページの期待値上位 30 位以内に，ユーザが閲覧したページが含まれているならば，推論結果が実データを予測したと考えることにする．実際には，ユーザが閲覧したページのうち，期待値上位 30 位以内に含まれるものの割合を予測率として計算し，これを評価する．例えば，推論用のデータの数を 2 個として推論を行い，結果，期待値上位 30 位以内には残りの証拠データのうち 6 個が含まれていたとすると，予測率は，

$$\frac{\text{(上位 30 位に含まれる推論用データ以外の証拠データの数)}}{\text{(全証拠データ数} \quad - \quad \text{推論用データ数)}} = \frac{6}{8} = 0.75$$

と求められる．

以上のことを踏まえて，実験は以下のような手順で行った．

1. ユーザが閲覧したページの中から推論に用いるページを選び，推論用データとする．
2. 推論用データから各ページの閲覧確率推論を行い，ページ閲覧回数の期待値を求める．
3. 期待値上位 30 位以内に，推論に用いたページ以外のユーザが閲覧したページがいくつ含まれるか数えあげる．
4. 3 で数えあげた値を，推論に用いたページ以外のユーザが閲覧したページの数で割り，予測率を求める．
5. 推論に用いるページを 1 つずつ増やして 1 から 4 の処理を繰り返し，推論用データの数ごとに予測率を求める．これを推論用データの数が 9 個になるまで行う．

4.3 結果

表 1: 評価結果

データ数	予測率平均値	標準偏差
0	0.387	0.217
1	0.707	0.158
2	0.679	0.162
3	0.683	0.166
4	0.663	0.176
5	0.665	0.186
6	0.676	0.217
7	0.667	0.238
8	0.647	0.310
9	0.647	0.478

予測率を，102 人のユーザすべてについて求め，各推論用のデータの数ごとに適合率の平均値と標準偏差を求めた．結果を表 1 に示す．データ数が 0 の項目は，推論用にデータを与えなかった場合の予測率である．評価結果から，なにも推論データを与えないときと比べて，データによる推論を行うことで予測率が高くなることがわかる．また，データによる推論を行った場合では，推論データの数によらず，予測率がほぼ 65 %を超えることがわかる．これらのことから，データからの推論結果が有効に使えることがわかる．

標準偏差がデータ数が多くなると大きくなる傾向にあるが，これはデータ数が増えると，その分，推論に使用しないユーザが閲覧したページが減り，取りうる予測値が制限されるためである．例えば，データ数が 9 個ある場合には，推論できるページは一つしかなく，予測値は 1 か 0 のどちらかしか取れなくなっている．

5. まとめ

本稿では，ベイジアンネットワークを用いた大規模な Web サイトを対象とした推薦システムの開発について報告した．さらに，実験的評価によりその有効性を示した．今後は，他の手法を用いた推薦システムや，MWST アルゴリズム以外の構造学習アルゴリズムとの性能比較などを行っていく予定である．

参考文献

[1] 繁桝算男, 植野真臣, 本村陽一, “ベイジアン・ネットワーク概要”, 培風館, 2006.

[2] Chow, C. K. and Liu, C. N., “Approximating discrete probability distributions with dependence trees”, IEEE Transactions on Information Theory, IT-14, pp.462-467, 1968.

[3] Cooper, G. F. and Herskovits, E., “A Bayesian methods for the Inducation of probabilistic networks from data”, Machine Learning, 9, pp.309-347, 1992.

[4] Schwarz, G., “Estimating the dimension of a model”, Annals of Statistics, 6, pp.461-464, 1978.

[5] Bozdogan, H., “Model selection and Akaike’s information criterion(AIC): The general theory and its analytical extensions”, Psychometrika, 52, pp.345-370, 1987.

[6] Cheng, J. and Greiner, R., “Comparing Bayesian Network Classifers”, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence,1999.

[7] Cheng, J., Hatzis, C., Hayashi, H., Krogel, M., Morishita, S., Page, D., and Sese, J., “KDDD cup 2001 report, ACM SIGKDD Explorations”, Vol. 3, No. 2, 2002.

[8] Pearl, J., “Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference”, Morgan Kaufmann Publishers, 1988.