社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (37)
2008－BIO－15 (37)
2008／12／18

# マルチベイジアンネットワークを用いたWebページのリコメンデーションシステム

左 毅, 上島康孝, 北 栄輔

名古屋大学大学院情報科学研究科

# Recommendation System of Web Page by using Multi Bayesian Network

Y. Zuo, K. Kamijima and E. Kita

Graduate School of Information Sciences, Nagoya University

本研究では，ブログ記事のURLを入力し，そのURLが示すブログと関連の強い別のブログを表示するためのWebページの推薦アルゴリズムについて述べる．方法(1)では，探索元のブログのキーワードを共有するブログを推薦し，それらのブログの並べ替えにはベイジアンネットワークを用いる．さらに，方法(2)では，探索元のブログ本文から形態素解析でキーワードを検索し，それを追加してブログの順位付けを行う．最後に簡単なテスト問題にアルゴリズムを適用する．

This paper describes the recommendation algorithms of blog pages which are strongly related to a user-specified blog page. Since each blog has some keywords, the recommended pages are selected from them which have the same keywords as the user-specified page. The recommended pages are sorted according to the Bayesian network. This is named as the Algorithm (1). In the Algorithm (2), keywords are selected from the body text of the user-specified blog by using a morphological analysis program. The keywords, in addition to the other keywords, are used for sorting the pages. Finally, the algorithms are applied to a simple example.

## 1 緒論

近年，インターネットの普及につれて，多くの人が個人のブログを掲示している．この中には，人気タレントや政治家，その他著名人などもあり，ブログアクセス数も着実に増加している．多数のブログから望みのWebページを得るために検索エンジンが利用される．しかし，既存の検索エンジンをブログ検索に用いる場合いくつかの問題がある．その一つは，既存の検索エンジンでは，ブログ記事のURLを入力し，そのURLが示すブログの内容と関連の強い他のブログを表示できないことである．

本研究では，上記の問題点を解決するためにベイジアンネットワークを用いる方法を提案する．あるキーワードで検索した結果，特定のブログのWebページが検索されたとき，このブログとキーワードの間にベイジアンネットワーク[1][2]を構築する．これを多数のブログについて実施した結果，異なるブログが同じキーワードで検索された場合，そのキーワードを介して異なるブログを結びつけることができる．このようにして，ブログに対して関連する別のブログを推薦する．これを方法1とする．さらに，ブログに対して既に設定されたキーワードだけでなく，ブログ本文の形態素解析からキーワードを抽出し，これを既に設定されたキーワードと組み合わせて利用することで方法1の性能を改善する方法である．本研究では，この2つの方法について述べる．

## 2 ブログ推薦方法

### (1) 方法1

各ブログ記事を親ノード，その記事について設定されたキーワードを子ノードとしてネットワークを構築する．他ブログ記事と2つ以上のキーワードを共有する場合，このようなブログ群は同じSenior集合とする．Senior集合に含まれるブログ記事は関連性があると言え，さらに，共有するキーワードが多いほど記事間の関連性が高い．このことを利用して検索を行う．

このアルゴリズムを簡単な例で述べる．同じSenior集合に含まれる2つのブログ記事BLOG1とBLOG2を考える．ここで，BLOG1に含まれて，BLOG2に含まれないキーワードXを考える．Xから検索を行うと，まずBLOG1が検索できる．そして，構築されたSenior集合からXを含まないBLOG2も検索できることになる．このようにして検索された全ブログ記事は，ベイジアンネットワークの条件付き確率によって順位付けされる．

### (2) 方法2

方法2と方法1の手法はほぼ同じである．方法1ではブログに既に設定されているキーワードだけを用いる．これに対して，方法2では検索の元になるブログの本文の形態素解析から得られたキーワードをもとのキーワード群に追加する．これから後の処理は方法1と同じである．

## 3 ブログのランキングアルゴリズム

提案するアルゴリズムでは，ブログ記事とキーワードの関係がNaïve-Bayes構造[3]に従っているとする．ベイジアンネットワークはベイズ推定[1][2]に基づいており，条件付き確率は次式で与えられる．

$$P(C_i \mid X_1,...,X_n) = \frac{P(C_i, X_1,...,X_n)}{\sum_j P(C_j, X_1,...,X_n)} \quad (1)$$

$$P(C_i, X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i)) \quad (2)$$

ここで$C_i$は原因としての事象を，$X_i$はその原因によって起きると想定される事象を示す．$P(C_i \mid X_1,...,X_n)$は事象$X_1,...,X_n$が発生した下で、事象$C_i$が発生する条件付き確率である。$Parents(X_i)$は$X_i$の原因となる事象の集合を示す．

ブログ記事Blogとキーワード$\mathrm{Keyword}_i$の関連性は，関連ないことを0，関連あることを1とする確率変数で関連づける．式(1),(2)において$C_i = \mathrm{Blog}$，および$X_i = \mathrm{Keyword}_i$とし，

Blog＝1 となる確率を表すと次式を得る．

$$P(\text{Blog}=1|\text{Keyword}_1,...,\text{Keyword}_n) = \frac{P(\text{Keyword}_1|\text{Blog}=1)\cdots P(\text{Keyword}_n|\text{Blog}=1)}{\sum_{\text{Blog}\in\{0,1\}} P(\text{Keyword}_1|\text{Blog})\cdots P(\text{Keyword}_n|\text{Blog})} \quad (3)$$

ここで，式(3)は各キーワードによってブログ記事が出現する確率を意味する．そこで，式(3)の右辺の確率変数をベイズ推定で求め，この確率を評価値として記事をソーティングする．

続いて，式(3)の右辺の評価方法について述べる．あるキーワード $\text{Keyword}_i$ について，Senior 集合に含まれるブログのうちで $\text{Keyword}_i$ を含むブログが占める割合を重み $\omega$，$\text{Keyword}_i$ を含む全てのブログのうちで Senior 集合に含まれないブログの割合を重み $\omega'$ とすると，ブログ記事とキーワードの条件付確率は表 1 のようになる．

表 1　ブログ記事とキーワードの CPT

| $P(\text{Keyword}_i\|\text{Blog})$ | Conditional Probability |
|---|---|
| $P(0\|0)$ | $1-\omega_i'$ |
| $P(1\|0)$ | $\omega_i'$ |
| $P(0\|1)$ | $1-\omega_i$ |
| $P(1\|1)$ | $\omega_i$ |

表 1 を用いると，式(3)から式(4)を得る．

$$P(\text{Blog}=1|\text{Keyword}_1,...,\text{Keyword}_n) = \frac{\omega_1(1-\omega_2)\cdots(1-\omega_n)}{\omega_1(1-\omega_2)\cdots(1-\omega_n)+\omega_1'(1-\omega_2')\cdots(1-\omega_n')} \quad (3)$$

この式は，キーワード $\text{Keyword}_i$ に関連する確率変数 $\omega_i$ が大きいほど評価値が大きい．同時に，$\text{Keyword}_i$ 以外のキーワードに対する確率変数 $\omega_i$ が小さいほど評価値が大きい．これにより，この Senior 集合に $\text{Keyword}_i$ を含むブログ記事の割合が大きく，$\text{Keyword}_i$ 以外のキーワードの割合が小さい記事を上位に推薦することになるからである．$\omega'$ は各 Senior 集合間の評価指標で，小さいほど評価値が大きい．同時に，$\text{Keyword}_i$ 以外のキーワードに対して，$\omega'$ が大きいほど評価値が大きい．これは，この Senior 集合に $\text{Keyword}_i$ を含むブログ記事の数が他の Senior 集合より多く，$\text{Keyword}_i$ 以外のキーワードを含むブログ記事の数が他の Senior 集合より少ないという意味で，推薦する時このような Senior 集合にある記事が上位になる．

## 4　数値実験

実験のために表 2 のようにブログ記事のデータベースを設定する．このとき，S1={B1, B2, B3, B4, B5}, S2={B6, B7, B8}, S3={B9}, S4={B10}の 5 つの Senior 集合が考えられる．

### (1) 方法 1

キーワード K3 で検索を行うと，Senior 集合間に K3 の表現の強弱($\omega'$)によって S1＞S2＞S3 の順位ができる．そして，各キーワードの $\omega$ と $\omega'$ を計算して各ブログ記事のスコアを算出する．このスコアによって，X3 を含むブログ記事は B1＞B6＞B2＞B3＞B9＞B7 の順番で上位に推薦される．そして，X3 を含まないブログ記事は B5 ＞ B8 ＞ B4 の順番で下位に推薦される．

実は，S1＞S2＞S3 によって，推薦順位は B1＞B2＞B3＞B6 ＞ B7 ＞ B9 となるべきである．しかし，K3 を含むブログ記事 B2，B3 と B6 に対して，B2，B3 は B6 より K3 の表現がほぼ同じである．K3 以外のキーワードに対して，B2，B3 は B6 より強く反映されて B2 と B3 のほうは $\omega$ が大きい．検索語以外のキーワードの表現は式(5)の$(1-\omega)$で評価し，$\omega$ が大きいのほうはスコアが小さい．更に，各 Senior 集合間の評価 $\omega'$ によって，検索語以外のキーワードの表現は B6 が B2 と B3 より強くて，$\omega'$ が大きい．これは，式(5)の$(1-\omega')$で評価し，$\omega'$ が大きいのほうはスコアが大きい．上述のことをまとめて式(5)で総合スコアを算出し，B6＞B2＞B3 となる．同様に，B9＞B7 となる．従って，最後の順位は B1＞B6＞B2＞B3＞B9＞B7 となる．

### (2) 方法 2

あるブログ B の記事を形態素解析したところ，キーワード K3, K4, K11 が抽出されたとする．抽出されたキーワードを記事に与えられたキーワードともに検索語として検索を行う．検索結果のソーティングは方法 1 の評価手法と同じようになる．まず，二つキーワード K3 と K4 を含む B2 と B3 を B2＞B3 の順番で推薦される．次に，一つキーワードを含む B1, B4, B6, B7, B9, B10 は，B4＞B1＞B10＞B9＞B6＞B7 の順番で推薦される．最後に，抽出されたキーワードを含まない B5 と B8 を B5＞B8 の順番で推薦される．

表 2　ブログ記事のデータベース

| Blog | Keyword | | |
|---|---|---|---|
| B1 | K1 | K2 | K3 |
| B2 | K2 | K3 | K4 |
| B3 | K3 | K4 | K5 |
| B4 | K4 | K5 | K6 |
| B5 | K5 | K6 | K7 |
| B6 | K3 | K6 | K8 |
| B7 | K3 | K7 | K8 |
| B8 | K2 | K7 | K8 |
| B9 | K3 | K9 | K10 |
| B10 | K11 | K12 | K13 |

## 5　結論

本研究では，あるブログのページに対して，関連するブログページをランキング付けして提示するシステムを提案した．本提案するシステムでは，ブログページが共有するキーワードを用いてブログ記事間を関連づけ，ブログの Senior 集合を定義する．ブログとキーワードの関連付けにベイジアンネットワークを利用し，ベイジアンネットワークの条件付確率表を用いてブログページの評価値を算出する．評価値の大きさによってブログページをランキングする．簡単な解析例を用いて手法の適用法を示した．

## 参照文献


[1] 繁桝算男，本村陽一，植野真臣：ベイジアンネットワーク概説，培風館（2006）．

[2] Heckerman, D., Geiger, D., and Chickering, D. : Learning Bayesian Networks : The Combination of Knowledge and Statistical Data, Machine Learning, Vol. 20, pp. 197-243 (1995).

[3] Cheng, J. and Greiner, R. : Learning Bayesian Belief Network Classifiers: Algorithms and System (2001).