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

2008－MPS－72 (38)
2008－BIO－15 (38)
2008／12／18

ベイジアンネットワークと静的嗜好情報を用いた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 Bayesian Network and Static Information of Interests

K. Kamijima, Y. Zuo, 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．提案する方法では，キーワードとWebページの間に構築したベイジアンネットワークの情報とユーザーの静的嗜好情報を用いる．最後に簡単な例でアルゴリズムの動作について説明する．

This paper describes the recommendation system of web pages which is strongly related to the user-specified web page. The system utilizes the information of the Bayesian network between the web pages and the keywords and the static information users' interests. Finally, the system behavior is explained in simple examples.

1 緒論

近年，インターネットの普及につれて多くの人たちが個人でブログを掲示している．それと共に，人気タレントや政治家等の著名人もブログを開設しており，アクセス数の多い女性タレントは“ブログの女王”と呼ばれ，クリックの世界だけでなくモルタルの世界における存在感を高めるためにも重要となっている．このような状況を鑑み，インターネットサービスプロバイダ(ISP)の側としても，サイトの集客力を高め，滞在時間を向上してリピータを増やすために，ブログを効果的に利用する必要がある．そこで，本研究では，ブログ記述者に対して，新たに掲載したブログの内容と関連の強い他のブログを提示するレコメンドシステムについて提案する．このようなシステムを効果的に用いることができれば，関連の強いブログ同士が相互リンクを張ることで同一ドメイン内のアクセスを互いに高め，集客力や滞在時間を向上できると考えられる．

提案手法では，ベイジアンネットワークと静的嗜好情報を利用する．ここで，静的嗜好情報とは，与えられたキーワード群からユーザー自身が選択した興味があるキーワード群のことである．提案手法のプロセスは以下のようになる．最初に，ブログとキーワードの間に構築したベイジアンネットワークの条件付確率表に基づいてブログの関連性の強さをランキングする．続いて，それらのブログに共通するキーワード分類をリストアップしてユーザーに提示する．ユーザーは，提示されたキーワード分類群について自身の興味の度合いを入力する．システムは入力された興味の度合いに応じて改めてブログをランキングし直し，その結果を提示する．本論文では，システムの理論的背景とシステム構成について述べる．

2 提案するシステム

2.1 システムの処理の流れ

提案するシステムの処理を箇条書きにすると以下のようになる．

1) ブログのWebページB0を入力する．
2) そのブログに著者が記述したキーワードをもとに関連が強いと思われるブログの集合BGを決定する．
3) B0,BGに共通するキーワード群をリストアップする．
4) リストアップされたキーワード群からキーワード分類群を構成する．
5) キーワード分類群のうち，関連の強い上位3つの分類群と関連性の弱い分類群からランダムに選ばれた2つをユーザーに提示する．
6) ユーザーはこれらの分類群についての重要度の強弱を入力し，検索ボタンをクリックする．これが静的嗜好情報の入力nあたる．
7) BGに含まれるブログを静的嗜好情報に基づいて並べ直してユーザーに提示する．

この方法で重要となるのは，関連するブログの集合BGを構成する方法と，静的嗜好情報に基づいて並び替えを行う方法の2点である．以下にこれらの点について説明する．

2.2 関連するブログの検索方法

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集合を構成する．Senior集合に含まれるブログ記事は関連性があると言え，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関連するブログを検索する．

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[図]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[山図]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，$\mathrm{Blog} = 1$ となる確率を表すと次式を得る．

$$P(\mathrm{Blog}=1 \mid \mathrm{Keyword}_1,\ldots,\mathrm{Keyword}_n) = \frac{P(\mathrm{Keyword}_1 \mid \mathrm{Blog}=1)\cdots P(\mathrm{Keyword}_n \mid \mathrm{Blog}=1)}{\sum_{\mathrm{Blog}\in\{0,1\}} P(\mathrm{Keyword}_1 \mid \mathrm{Blog})\cdots P(\mathrm{Keyword}_n \mid \mathrm{Blog})} \quad (3)$$

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 $\mathrm{Keyword}_i$ について，Senior 集合に含まれるブログのうちで $\mathrm{Keyword}_i$ を含むブログが占める割合を重み ω，$\mathrm{Keyword}_i$ 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 ω' 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

Keyword	Blog	P
0	0	$1-\omega'$
1	0	ω'
0	1	$1-\omega$
1	1	ω

表 1 を用いると，式(3)から式(4)を得る．

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

この式は，キーワード $\mathrm{Keyword}_i$ に関連する確率変数 ω_i が大きいほど評価値が大きい．同時に，$\mathrm{Keyword}_i$ 以外のキーワードに対する確率変数 ω_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 $\mathrm{Keyword}_i$ を含むブログ記事の割合が大きく，$\mathrm{Keyword}_i$ 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．ω' 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，$\mathrm{Keyword}_i$ 以外のキーワードに対して，ω' が大きいほど評価値が大きい．これは，この Senior 集合に $\mathrm{Keyword}_i$ を含むブログ記事の数が他の Senior 集合より多く，$\mathrm{Keyword}_i$ 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．

2.3 静的嗜好情報による並び替え

データベースから全てのブログのキーワードを抽出し，tf-idf 法で各キーワードの評価値を算出しておく．そして，入力されたブログ記事に与えられたキーワードを評価値の高い順に並べる．その中で，評価値の高いもののうちの上位から 3 つと，それ以下のうちからランダムに選択された 2 つをユーザーに提示する．ユーザーは自分が好きなキーワードを選択し，2.2 で検索された結果に基づいて再ソーティングを行う．選択したキーワードを含む記事はそのキーワードの評価値を得点として加算する．このようにして，得点が 0 以外のブログを抽出して，高い順でユーザーに推薦する．この時，得点が同じの場合に元の順番で保持して並べる．最後，残るブログ（得点が 0 である）は 2.2 で求めた順番を保持し，ユーザーに推薦する．

3　解析例

実験のために，ブログ記事のデータベースを表 2 のように設定する．ここでは，次の 5 つの Senior 集合が考えられる．

S1={B1, B2, B3, B4, B5},
S2={B6, B7, B8}, S3={B9}, S4={B10}

続いて tf-idf 法で評価値を求めると，各キーワードは次のようにランキングされる．

K2, K4, K5, K6, K7, K8, K3, K1, K9, K10, K11, K12, K13

続いて，あるブログ

B={K2, K3, K4, K5, K6, K9, K10, K11, K13}

を入力して検索を行う．2.2 のアルゴリズムに従えば，各ブログのランクは次のようになる．

B3 > B4 > B2 > > B5 > B7 > B10 > B8 > B1 > B9 > B6

次は，キーワード群に上位から K2, K4, K5 と下位からランダムで K10, K13 を選択し，ユーザーに提示する．ここで，K2, K4, K13 を選択すると想定され，ブログの推薦順を再ソーティングする．得点によって，次のようになる．

B2 > B1 = B3 = B4 = B8 > B10 > B5 = B6 =B7 = B9

最後は 2.2 節の手法により次のようになる．

B2 > B3 > B4 > B8 > B1 > B10 > B5 > B7 > B9 > B10

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

4　まとめ

本研究では，あるブログに対して，関連があるブログをランキング付けして提示するシステムを提案した．提案システムでは，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を関連づけ，関連性のある候補ブログの集合を求める．続いて，ユーザーが入力した静的嗜好情報に基づいて候補ブログを並べ替える．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用いる．最後に，簡単な例でアルゴリズムを説明した．

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).