確率的インデックスによる事例ベース検索

阿部仁紀　藤原祥隆　前田康成　吉田秀樹

北見工業大学工学部

1．はじめに

我々の研究室では，対面教育における学習者の満足度向上と教師の負担軽減を目的とする対面教育支援システムの研究を行っている[1]．図1にその構成を示す．学習者の質問に回答することを目的とする支援要求推定の実現のため，質問内容とその解決方法を対にした事例の集合（以下，事例ベース）から，確率的インデックス法を用いて適切な回答を検索する事例ベース検索を確立した．

本稿では，確率的インデックス法を用いた事例ベース検索およびその評価について報告する．

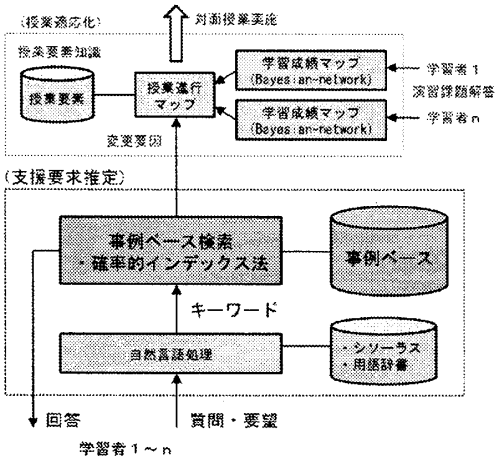

図1　対面教育支援システムの構成

2．事例ベース検索の概要

本事例ベース検索は，与えられたキーワードと概念番号を手がかりに，事例ベースの中から適切な事例を検索し，その事例の解決方法を回答として返すものである．ここでキーワードとは，質問の内容を特徴付ける文字列であり，概念番号とは，類似するキーワードを一つのグループとしてまとめ識別番号を与えたものである．例えば，「画像，絵，図」を一つのグループとし，識別番号 S10 を与えるようなものである．これらの識別番号とキーワードの対応を記録したものが図1中のシソーラスである．概念番号の利用は，事例ベース検索の性能向上を目的とするものである．以後，キーワードと概念番号をまとめてキーワードと記述する．

3．確率的インデックス法

本事例ベース検索の特徴は，確率的インデックス法を用いた検索機構にある．本インデックス法は，事例ベースをもとに Bayesian network として構成する．キーワードと事例をそれぞれノードで表し，キーワードノードと関連のある事例ノードの間にリンクを形成する．各リンクには，キーワードノードと事例ノードの間の関連の強さを反映した条件付確率の値を設定する．各ノードは，{存在（y），非存在(n)}の二つの状態をもつ．

Case-base retrieval using a probabilistic indexing method
Satoki Abe, Yoshitaka Fujiwara, Yasunari Maeda,
and Hideki Yoshida,
Dept. of Computer Sciences, Kitami Institute of Technology

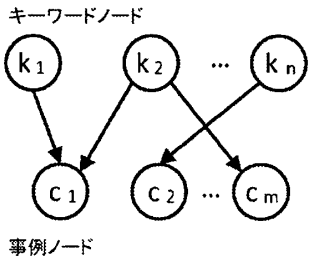

図2　インデックスの構成例

全事例ノードの集合Cと全キーワードノードの集合Kを下記のように定義する．

$$C=\{c_1,c_2,\cdots,c_m\} \qquad (m=|C|)$$

$$K=\{k_1,k_2,\cdots,k_n\} \qquad (n=|K|)$$

Bayesian network は図2のようになる．ここである事例ノードc_j $(\in C)$に注目する．事例ノードc_jとリンクをもつキーワードノードの集合K_j'を下記のように定義する．

$$K_j'=\left\{k_{j1}',k_{j2}',\cdots,k_{jn_j}'\right\}\subseteq K \qquad (n_j=|K_j'|)$$

事例ノードc_jは一般に複数のキーワードノードk_{jr}' $(\in K_j'.\ r=1,2,\cdots,n_j)$とリンクをもち，リンクには条件付確率 $P\left(c_j=y\middle|k_{j1}'\in\{y,n\},k_{j2}'\in\{y,n\},\cdots,k_{jn_j}'\in\{y,n\}\right)$ を$k_{j1}',k_{j2}',\cdots,k_{jn_j}'$の状態 y と状態 n の組み合わせについて与えなければならない．キーワードノードがn個あるとすると求める条件付確率は2^n個あり，nが増えるたびに求める条件付確率は指数オーダーで増加する．そのため，この条件付確率を求めるのは容易ではない．そこで本インデックス法では"Noisy-Or"の規則[2]を適用することにより，$P\left(c_j=y\middle|k_{j1}',k_{j2}',\cdots,k_{jn_j}'\right)$を下記の式(1)と表すことができる．ただし，$I(k_{jr}')$は$k_{jr}'$に対応する$K$中のキーワードノード$k_i$の添え字$i$を与える関数とする．

$$P\left(c_j=y\middle|k_{j1}',k_{j2}',\cdots,k_{jn_j}'\right)\fallingdotseq 1-\Pi_{I(k_{jr}')\in Y}\,q_{I(k_{jr}')j} \qquad (1)$$

ただし，Yはc_jとリンクをもち，状態 y であるキーワードノードk_iの添え字iの集合を表し，

$$q_{I(k_{jr}')j}=P\left(c_j=n\middle|k_{jr}'=y\right) \qquad (r=1,2,\cdots,n_j) \qquad (2)$$

とする．また，状態 y のキーワードが存在しない場合は，式(1)の値を0とする．

本インデックス法では，式(1)を「事例c_jが質問に対応する事例である確率（以下，確からしさ）」とする．また式(2)をノード間の条件付確率とし「質問から得られたキーワードk_{jr}'が存在するとき，事例c_jが質問に対応する事例ではない確率」とする．

与えられたキーワードをもとに事例ベース検索を実行すると，以下の手順をとる．まず与えられたキーワードに対応するキーワードノードを状態 y とする．次に，状態 y であるキーワードノードと事例ノードの間の条件付確率である式(2)の値を用いて，各事例ノードの確からしさである式

(1)を計算する．計算された値から”確からしい”事例の解決方法の内容を返す．例えば最も値が大きい事例を候補としたり，値に順位をつけて値が 0.3 以上の事例を複数候補として提示し，最も適切な事例を選択させるなどのバリエーションが考えられる．

4．本インデックス法の評価

4．1．登録事例を使用した性能評価実験

我々の研究室が開発した e-ラーニングシステム[3]に関する質問・回答を対象とした 43 個の事例から事例ベースを作成し，本インデックス法の評価を行った．検索で使用するキーワードは，事例の質問に含まれる名詞とした．ノード間の条件付確率を以下のように設定した．

まず，3 章で定義した全事例ノードの集合Cと全キーワードノードの集合Kのもとで，あるキーワードノード$k_i\ (\in K)$に注目する．キーワードノードk_iとリンクをもつ事例ノードの集合C_i'を下記のように定義する．

$$C_i' = \{c_{i1}', c_{i2}', \cdots, c_{im_i}'\} \subseteq C \qquad (m_i = |C_i'|)$$

キーワードノードk_iとリンクをもつ事例ノード$c_{ir}'(\in C_i'.\ r = 1,2,\cdots,m_i)$が全部で$m_i$個あるとき，$P(c_{ir}' = y|k_i = y)$を式(3)とした．

$$P(c_{ir}' = y|k_i = y) = \frac{1}{m_i} \qquad (r = 1,2,\cdots,m_i) \qquad (3)$$

キーワードノードk_iと事例ノードc_{ir}'の間の条件付確率は式(2)と式(3)より，下記の式(4)とした．ただし，$J(c_{ir}')$はc_{ir}'に対応するC中の事例ノードc_jの添え字jを与える関数とする．

$$q_{ij\,(c_{ir}')} = 1 - \frac{1}{m_i} \qquad (r = 1,2,\cdots,m_i) \qquad (4)$$

次に上記の 43 個の事例を使用し，各事例に対する質問の発生頻度にばらつきのあるユーザの質問モデルを作成した（質問の発生頻度は，最高が 5.5%，最低が 0.3%，平均が 2.3%）．そしてこのモデルにより 100,000 回の質問を生成し，質問に対する回答を検索して，得られた回答がどの程度正しいかを評価した．評価尺度としては，質問に対する正しい回答が第一位に検索される割合（以下，検索率）を使用した．表１の「初期インデックス」にこの場合の検索率を示す．

さらに，上記モデルから生成された 100,000 回の質問と回答の履歴データベースを使用して，式(4)のノード間の条件付確率を以下の方法でチューニングした．

事例c_{ir}'が履歴データベースに出現した回数を$M_{ir}\ (r = 1,2,\cdots,m_i)$とする．このとき，キーワード$k_i$と事例$c_{ir}'$に対応するキーワードノードと事例ノードの間の条件付確率は下記の式(5)とした．

$$P(c_{ir}' = y|k_i = y) = \frac{M_{ir}}{T_i} \qquad (r = 1,2,\cdots,m_i)$$

$$q_{ij\,(c_{ir}')} = 1 - \frac{M_{ir}}{T_i} \qquad (5)$$

ただし，$T_i = \sum_{r=1}^{m_i} M_{ir}$

式(5)によりキーワードノードと事例ノードの間の新たな条件付確率を求め，値を更新した．

そして，チューニングされた条件付確率のもとで，上記のユーザ質問モデルに従って，新たに 100,000 回の質問を生成し，質問に対する回答を検索し，検索率を求めた．表１の「チューニング後のインデックス」にこの場合の検索率を示す．表１より，履歴データベースを使用したチューニングを行うことにより，検索率の改善が期待できることがわかった．

表 1　本インデックスの性能評価実験(1)

	初期インデックス	チューニング後のインデックス
検索率	72.3%	96.9%

4．2．実際の運用を想定した性能評価実験

4.1.の性能評価実験で使用した 43 個の事例が登録された事例ベースを用いて，複数の実験協力者より得られた計 40 個の質問で検索を行い，質問ごとに下記のいずれかの評価を与えた．

成功：知識あり　適切な回答を第三位以内に提示した．
成功：知識なし　事例ベースに適切な事例がなく，かつ「該当する情報はありません」と回答した．
失敗：知識あり　事例ベースに適切な事例があるにも関わらず，回答を提示できなかった．
失敗：知識なし　事例ベースに適切な事例がないにも関わらず，不適切な回答を提示した．

評価結果を表２に示す．

表 2　本インデックスの性能評価実験(2)

成功	知識あり	5 (12.5%)
	知識なし	3 (7.5%)
	計	8 (20.0%)
失敗	知識あり	5 (12.5%)
	知識なし	27 (67.5%)
	計	32 (80.0%)
合計		40 (100.0%)

表２より，成功率 20%とあまり高い値は得られなかった．質問が事例ベースにないケース（知識なし）が 75%もあり，事例ベースに登録された事例が十分ではないことがわかる．成功率を上げるためには，多くの事例を事例ベースに用意する必要があるだろう．またこの実験では，事例の”確からしさ”が 0 でないものをすべて回答として提示している．”確からしさ”が低い（例えば 0.3 以下の）場合は，「該当する情報はありません」と回答することで，「失敗：知識なし」の割合を下げることができるだろう．「成功：知識あり」は，回答が第三位以内に提示された場合と定義したが，第三位以内と限定しなかった場合，成功率は 30%に上がり，「失敗：知識あり」の割合は 2.5%に下がった．「失敗：知識あり」については，質問が事例ベースに存在するにも関わらず，なぜ正しい回答が得られなかったのか，調べる必要がある．

5．おわりに

本稿では，確率的インデックス法を用いた事例ベース検索およびその評価について報告した．今後は，性能評価実験で得られた結果をもとに，更なる検索率の向上を求め検討を進める予定である．

参考文献

[1] 福島潤一郎, 藤原祥隆, 前田康成, ”確率的推論を基礎とする学習者マップを利用した対面教育適応化法”, FIT2007 第 6 回情報科学技術フォーラム, pp. 589-590, Sep 2007
[2] Finn V. Jensen, *Bayesian Networks and Decision Graphs (Statistics for Engineering and Information Science)*, 284pp, SPRINGER, BERLIN, 2001,
[3] 鈴木智樹, 藤原祥隆, 岡田信一郎, 吉田秀樹, “ユーザ適応化 e-Learning システム KUSEL の設計”, 情報処理学会研究報告 2004-ICS-135, vol. 2004 no. 19, pp. 169-174, Mar 2004