失敗からの知識獲得とテキスト分類に基づくインターネットからの情報収集

堀井則彰[†] 上原邦昭[†,††]

ネットワークを介して入手できる情報は膨大であり，ユーザがすべての情報に目を通すことは不可能である．このため，多くの情報の中からユーザにとって必要な情報のみを抽出する，情報検索の開発が望まれている．本論文では，ベイジアンネットワークを導入したクラスタリングを用いてWebページの階層木を構築する手法を提案する．さらに，内容が類似しているWebページは階層木中で近接しているという仮定に基づいて，ユーザの関心を満たすWebページが抽出される．このとき，興味のないWebページが検索された場合には,「失敗からの知識獲得」に基づいて漸増的にベイジアンネットワークが構成される．このため，ユーザの関心に応じてWebページの階層木が更新されるようになっている．

Information Gathering with Text Categorization and Knowledge Acquisition by Failure

Noriaki Horii[†] and Kuniaki Uehara[†,††]

A spread of network and information service easily let us obtain a vast amount of useful information through the network. Information retrieval becomes an important tool for gathering desirable information. In this paper, we will propose a personalized information gathering system that constructs a hierarchical tree of Web pages represented into vector models by a Bayesian network. The system retrieves relevant Web pages based on the assumption that similar Web pages are clustered closely. When the system encounters a retrieval failure, which is that irrelevant Web pages have been retrieved, the Bayesian network is updated by "knowledge acquisition by failure." Therefore, the hierarchical tree varies with the user's preference, and the system can retrieve relevant information based on interests of an individual user.

1. はじめに

ネットワークや情報サービスの普及により，遠隔地のさまざまな情報を容易に入手できるようになった．このため，大量の情報から必要なデータのみを検索するメカニズムが必要とされている．しかしながら，現在WWW上で実用化されている情報検索エンジンは不特定多数を対象とした汎用的なシステムが多く，個人ユーザの関心に基づく検索は困難である．なかには，ソーシャルフィルタリングを用いてWebページを紹介するシステム☆もあるが，他人が頻繁に見るWebページが特定ユーザの興味を満たしているとは限らず，必ずしも個人ユーザの関心に基づく検索エンジンになっているとはいえない．

本論文では，ベイジアンネットワーク[10]を用いた個人ユーザの関心に基づく情報収集について述べる．本システムでは，問題領域に関する知識やユーザの関心を領域知識としている．一般に，領域知識としてはシソーラスが使用されるが，ユーザの嗜好や問題領域に関する概念は時間とともに変化するため，静的なシソーラスが動的な環境でも有効に働くとは限らない．また，領域知識としてシソーラスを導入したとしても，新規の分野に関するデータの保守や準備には専門家が必要となる．このため，本システムでは，自動的なデータの更新が可能なベイジアンネットワークを用いて領域知識を表現している．ベイジアンネットワークの各ノードは問題領域の単語に対応し，単語間を結ぶアークは単語の共起出現確率を表している．共起出現

† 神戸大学工学部情報知能工学科
Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University
†† 神戸大学都市安全研究センター
Research Center of Urban Safety and Security, Kobe University

☆ たとえば，http://okiraku.navi.ntt.co.jp などがある．

確率を利用すれば，情報検索時には同義語となる単語の追加，あるいはノイズとなる単語の削除など，より個人ユーザに特化した検索が可能となる．

本システムは，ユーザが以前に発見した興味のあるWebページや論文（desired textと呼ぶ）を目標概念とし，目標概念に類似した概念の獲得，つまりユーザの関心を満たすWebページ（正解ページと呼ぶ）の検索を目的としている．具体的には，ユーザがWWW上のブラウジングで発見した興味のあるWebページをdesired textとしてブックマークに入れておけば，desired textのURLを指定するだけで正解ページを検索することができるようにしている．

一方，検索対象となるWebページは，ベイジアンネットワークを利用したクラスタリングによって階層木に分類される．言い換えると，階層木の近接した葉には類似したWebページが配置されるようになる．したがって，正解ページを検索する場合には，まずdesired textを階層木に分類し，desired textの近くに分類されたWebページを正解ページとして提示している．

もし検索が失敗すれば,「失敗からの知識獲得」に基づいてベイジアンネットワークが洗練される．ここで，検索の失敗とは，ユーザが関心を持たないWebページ（不正解ページと呼ぶ）がdesired textの近くに分類されることである．これは次の3点が原因となっている．

- Webページ間の意味的な違いを認識していない場合
- Webページ間で共通している話題を誤認識している場合
- Webページ中で重要な記述を誤認識している場合

これらの失敗をトリガとして，それぞれの失敗に基づいた個別の知識獲得が行われる．具体的には，新たな単語の獲得，単語の共起出現確率の変更，ネットワークのトポロジの変化により，ユーザの「個人的な」問題領域に関する知識が構築される．これは，不正解ページが遠ざけられ，desired textの近傍には正解ページのみが集まるようにベイジアンネットワークを構成することに相当している．ベイジアンネットワークの更新が終わると，Webページの再クラスタリングが行われ，より多くの正解ページが抽出されるようになる．

2. 本システムの概要

2.1 ベイジアンネットワークの構築と推論

本システムをはじめて利用する際には，各ユーザごとの領域知識を表すベイジアンネットワークが作成される．これを初期ベイジアンネットワークと呼ぶ．初期ベイジアンネットワークは以下の手順に従って構築される．

(1) 検索エンジンにキーワードを与え，キーワードを含むWebページ集合を求める．

(2) Webページ集合に出現している単語のTFIDF☆を計算する．TFIDFの値が上位にランクされている単語を，問題領域で使用される単語として単語集合Wordに入れる．なお，この時点でWordから機能語を削除する．

(3) TFIDFが最大の単語をルートとし，Wordから削除する．次に，ルートの単語とWordに含まれている単語の相互情報量$\log(p(x\&y)/(p(x)p(y)))$を計算し，しきい値以上である単語をルートの子ノードとし，Wordから削除する．また，しきい値以下である単語はWordに残し，何もしない．ただし，$p(x\&y)$は単語xとyが同じ文中で共起して出現する確率，$p(x)$および$p(y)$は単語xおよびyが出現する確率である．なお，アークに付随した確率は$p(x\&y)$を示している．

(4) 子ノードを新たな親ノードとし，上記と同様にして，再び子ノードとなる単語をWordから発見していく．この操作をWordが空になるまで繰り返し，ベイジアンネットワークを構築する．

以上の手順により，木構造のベイジアンネットワークがトップダウン的に形成される．

構築されたベイジアンネットワークで，いくつかの単語を証拠変数とすれば，ある単語（$Word$）が証拠変数と共起して出現する確率（$Prob$）は，式(1)に基づいて推論することができる．

$$Prob = PaProb \times CoProb + (1 - PaProb) \times UniqProb \quad (1)$$

ただし，$PaProb$は$Word$の親ノードに属する単語（$Parent_{Word}$）の出現確率，$CoProb$は$Word$と$Parent_{Word}$の共起出現確率，$UniqProb$は$Parent_{Word}$が出現しない場合の$Word$の出現確率を表している．たとえば，単語“workshop”が“professor”と共起して出現する確率を推論する（**図1**参照）．このとき，“professor”が証拠変数であるという情報は，ベイジアンネットワークのすべてのノードに伝搬する．このため，“univ”の出現確率は$0.19(1\times0.19+(1-1)\times+0.24)$となり，“workshop”

☆ 単語の出現頻度（Term Frequency）と単語を含む文書の分散（Inverse Document Frequency）を掛けあわせたものをTFIDFという．

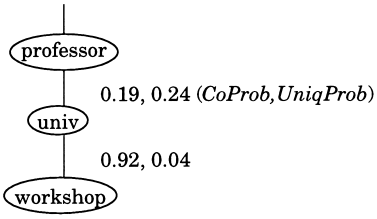


図 1 ベイジアンネットワークでの推論
Fig. 1 Inference by using a Bayesian network.

の出現確率は $0.21(0.19\times0.92+(1-0.19)\times0.04)$ となる．この結果，“workshop” は 0.21 の確率で “professor” と共起して出現すると推論される．

2.2 Web ページのデータ表現

一般に，情報検索の分野ではテキスト文書をデータ表現するために，単語の出現頻度を用いることが多い．単純に出現頻度のみを用いると，データ表現に要する情報量が非常に大きくなってしまう．さらに，Web ページには多くのノイズ，つまり内容とは関係のない単語が多く含まれている．このため，Web ページの内容を単語の出現頻度のみで正確に表現することは困難である．また，人手による Web ページの索引づけには問題領域に関する専門家を必要とし，その負担も大きなものとなってしまう．そこで，本システムでは重要語と呼ぶ概念を導入し，重要語の出現の有無を用いて Web ページのデータ表現を行っている．

重要語とは Web ページに含まれている重要な単語であり，TFIDF の値が上位にランクされている単語を重要語としている．Web ページを属性–値の対で記述する際には，重要語が含まれていれば属性の値を 1，含まれていなければ 0 としている．たとえば，次のようなテキスト文書を考える．

> **Conventional information retrieval is very closely related to information filtering in that they both have the goal of ret- rieving information relevant to what a user wants, while minimizing the amount of irrelevant information retrieved.**

仮に “extraction”，“filtering”，“profile”，“text” が重要語であるとすれば，このテキスト文書は (0, 1, 0, 0) (=(extraction, filtering, profile, text)) として表現される．

2.3 ノイズとなる重要語の除去

Web ページには，研究者の研究テーマの一覧やリンク集のような，重要な単語が箇条書きされているページがある．このようなページには重要語が多く含まれているが，個々の重要語が指し示す問題領域についてはあまり記述されていない．一般に，ある重要語が出現していたとしても，同一の問題領域に属する他の重要語が出現していなければ，その問題領域に関する情報はほとんど含まれていないと考えられる．また，文書中での重要語は，ランダムに出現するのではなく群を形成して出現している[3]．

一方，ベイジアンネットワークでは，頻繁に共起して出現している単語はアークで結ばれている．このため，意味的に関連があり，Web ページ中で群を形成している重要語は，ベイジアンネットワークでも近接して配置される．逆に，ノイズとなる重要語はベイジアンネットワーク中では孤立して配置されていると考えられる．したがって，Web ページをデータ表現する際には，ベイジアンネットワーク中で孤立している重要語の属性値を 0 としている．具体的には，以下の手順を用いてノイズとなる重要語を発見している．

(1) ベイジアンネットワークと Web ページに出現している重要語を Wordset に入れる．
(2) Wordset から重要語（Word）を 1 つ取り出し，Word が出現している文とその前後の文中に出現している重要語を Co-occur に入れる．たとえば，**図 2** の文書中で documents という単語が Word に含まれていれば，**図 3** から retrieval, feedback, query, term も Co-occur に入れられることになる．
(3) Co-occur に含まれている重要語を証拠変数，Word を質問変数とし，推論によって Word の出現確率を求める．推論された確率がしきい値以上であれば Word の属性値を 1，それ以外なら 0 とする．たとえば図 3 では，retrieval, feedback,

While this body of work is not necessarily focused exclusively on the retrieval problem, it demonstrates effectively how learning can be used to improve queries. Relevance feedback are a form of supervised learning where a user indicates which retrieved documents are relevant or irrelevant. These approaches have investigated techniques for automatic query reformulation based on user feedback, such as term reweighting and query expansion.

図 2 Web ページの文書
Fig. 2 Document in a Web page.

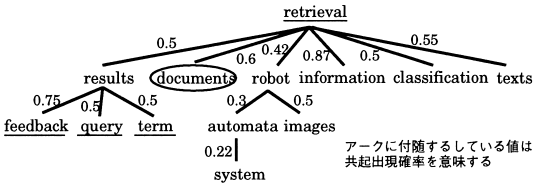


図 3 ベイジアンネットワーク
Fig. 3 Bayesian network.

query，termが証拠変数，documentsが質問変数となり，documentsの出現確率は0.584と推論される．しきい値が0.5の場合，documentsは重要語と評価されて属性値は1となる．

(4) Wordsetから Wordを削除し，Wordsetが空になるまで上記の操作を繰り返す．

2.4 同義語となる重要語の付加

Webページには同じ意味を表す場合でも，複数の異なった単語が使用されていることがある．たとえば，情報検索の分野では，“The system retrieves information.” “The system retrieves documents.” “The system retrieves texts.” はすべて同じ概念を表しており，“information”，“documents”，“texts” は同じ意味で使用されている．したがって，Webページをデータ表現する際には，同義語について考慮する必要がある．

一般に，2つの語の同義性を決定する指標として，意味論の分野では交換可能性という考え方が提案されている[7]．これは，2つの語が同義であるとは，それぞれの語が現れる文脈中で互いに他の語に入れ換えても，文脈が全体として何らかの意味で変わらない場合をいう．この考え方をベイジアンネットワークに適用すると，同義語となる重要語は同じ親ノードを共有していると考えられる．つまり，ベイジアンネットワークで兄弟関係にあるノードの重要語は，同義語の関係となっている可能性が高いことになる．したがって，本システムでは次の手順を用いて同義語を発見している（図2，図3参照）．

(1) Wordが Webページ中で出現している文とその前後の文中で出現している重要語を証拠変数とする．さらに，Wordも証拠変数とする．たとえば，図2で Wordに documentsが含まれていれば，documentsと retrieval，feedback，query，termを証拠変数とする．

(2) Wordと兄弟関係にある重要語を質問変数として，出現確率を求める．推論された確率がしきい値以上ならば，その重要語を Wordの同義語と見なして属性値を1，それ以外なら0とする．たとえば図3のベイジアンネットワークで Wordが documentsの場合には，results，robot，information，classification，textsが質問変数となり，それぞれの出現確率は1.0，0.42，0.88，0.49，0.55と推論される．しきい値が0.5である場合には，results，information，textsが documentsの同義語と評価されて属性値は1となる．

この操作により，シソーラスを用いずに同義語の情報をデータ表現に付加することができるようになっている．

2.5 Webページの検索

本システムでは，COBWEBアルゴリズム[4]に基づいて Webページを階層木にクラスタリングしている．また，構築された階層木では意味的に類似した Webページが近接して分類されている．言い換えると，階層木中で desired textの近くに分類されている Webページは，desired textと意味的に類似していることになる．したがって，それらのページを正解ページとして抽出している．具体的には，次の手順を用いて関心のある Webページを検索している．

(1) 入力された desired textを Webページの階層木へクラスタリングする．

(2) desired textの親となるノードを Nodesetへ入れる．

(3) Nodesetに含まれるノード（Node）の特徴ベクトルを作成する．Nodeの特徴ベクトルとは，属性を重要語とし，重要語が出現している確率を属性値とする属性–値の対で表現される．

(4) Nodeの特徴ベクトルと desired textのユークリッド距離を計算する．計算された距離がしきい値以内であれば，Nodeを Displayに入れ，Nodeの親ノードと子ノードを Nodesetに入れる．また，Nodeは Nodesetから削除する．

(5) 上記の操作を Nodesetが空になるまで繰り返し，最終的に Displayに含まれている Nodeの葉に分類されている Webページが，正解ページとしてユーザに提示される．

2.6 失敗からの知識獲得

本システムでは，検索の失敗をトリガとして，ユーザとの対話による「失敗からの知識獲得」により，漸増的にベイジアンネットワークを構築している．検索の失敗とは，不正解ページを提示してしまった場合である．ユーザから不正解ページが指摘されると，システムは指摘された Webページが正しく分類されるべきノード（負ノードと呼ぶ）を階層木中から発見する．このとき，階層木中での負ノードの位置に基づいて，検索の失敗原因を特定することができる．最終的に，検索の失敗に応じてベイジアンネットワークが更新される．これを「失敗からの知識獲得」と呼ぶ．

2.6.1 負ノードの選択

すでに述べたように，負ノードは不正解ページが本来属すべきノードである．このため，負ノードと不正解ページには何らかの類似性がある．また，負ノード

と正解ページにも何らかの類似性がある．したがって，候補となるノードのうち不正解ページと最も類似し，正解ページと最も類似していないノードを負ノードとしている．また，このような考え方に基づいて，以下の手順で負ノードを発見している．

(1) ノードの特徴ベクトルと desired text のベクトルのユークリッド距離がしきい値以上と判断されたノードを，負ノードの候補として Cnd に入れる．

(2) Cnd に含まれるすべてのノード N に対して，不正解ページとの類似度（Sim_{N}）と，正解ページとの相違度（$Diff_{\mathrm{N}}$）を計算する．Sim_{N} は，ノード N と不正解ページに共通している単語の TFIDF の総和（Com_n）と，共通していない単語の TFIDF の総和（$Uniq_n$）を用いて定義している．

$$Sim_{\mathrm{N}} = \log(Com_n/Uniq_n) \qquad (2)$$

この式は，不正解ページと意味的に類似した負ノードを見つけることを目的としたものである．したがって，類似度が高くなるほど Sim_{N} が大きくなるように定義している．

また，$Diff_{\mathrm{N}}$ はノード N と正解ページに共通している単語の TFIDF の総和（Com_p）と，共通していない単語の TFIDF の総和（$Uniq_p$）を用いて定義している．

$$Diff_{\mathrm{N}} = \log(Uniq_p/Com_p) \qquad (3)$$

この式は，正解ページと意味的に類似していない負ノードを見つけることを目的としている．したがって，類似度が低くなるほど $Diff_{\mathrm{N}}$ が大きくなるように定義している．

(3) 不正解ページの類似度と正解ページの相違度の和が最も高いノードが，負ノード（Neg）として選択される．

$$Neg = \arg\max_{c \in \mathrm{Cnd}} (Sim_c + Diff_c) \qquad (4)$$

不正解ページとできるだけ意味的に類似し，正解ページと類似していないノードを負ノードとして定義している．このため，Sim_{N} と $Diff_{\mathrm{N}}$ の和が最も高いものを Neg と定義している．

なお，$Uniq$ を計算するときの単語の数は Com の計算で用いた単語の数とし，TFIDF の上位から順に選択している．

2.6.2 検索の失敗の特定

負ノードが選択されると，負ノードと不正解ページ，正解ページの相対的な位置関係により，検索の失敗の原因が特定される．**図 4** の階層木では，システムがユーザに提示した Web ページを葉とするノードを黒丸で表している．また，白丸の葉にあたる Web ページはユーザに提示されなかったことを表している．さらに，黒丸の葉のうち，不正解ページ（irrelevant Web page）と明示されていないもの以外は正解ページとしている．このような状況で，検索の失敗としては次の 3 種類について考えている．

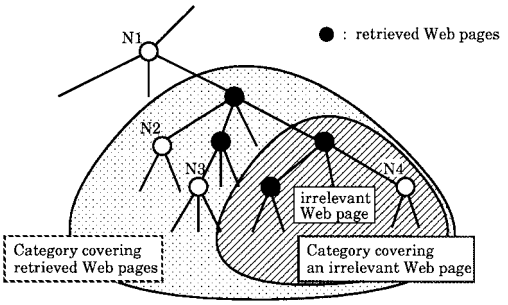


図 4 Web ページの階層木
Fig. 4 Hierarchically stored Web pages.

Web ページ間の意味的な違いの誤認識

不正解ページの属するノードあるいはその子孫以外に負ノードが位置している場合を考える．すなわち，N1，N2，N3 のいずれかが負ノードとして選択された場合である（図の斜線部以外にあるノード）．これは，不正解ページが内容と合致しないカテゴリに分類されていることを意味している．この誤分類は，ベイジアンネットワークに含まれている重要語が欠落しているため，Web ページの内容が誤認識されたことが原因と考えられる．

Web ページ間の共通話題の誤認識

また，正解ページの属するノードあるいはその子孫以外に負ノードが位置している場合を考える．すなわち，N1 が負ノードとして選択された場合である（図の点線部以外のノード）．これは，正解ページと不正解ページの問題領域が異なっていることを意味している．異なる問題領域にもかかわらず，Web ページが同じカテゴリに分類されるのは，偶然にも同じ重要語が使われていることが原因と考えられる．

Web ページでの重要な記述の誤認識

さらに，正解ページの属するノードあるいはその子孫に負ノードが位置している場合を考える．すなわち，N2，N3，N4 のいずれかが負ノードとして選択された場合である（図の点線部の中にあるノード）．これは，不正解ページと desired text の問題領域は同一であるが，共通している内容の重要度が異なっていることを意味している．このような場合，それらの Web ペー

ジは重要な内容に最も合致したサブカテゴリに分類されることが望ましい．

2.6.3 ベイジアンネットワークの更新

新たな重要語の追加

問題領域に関する重要語が欠落している場合，Webページ間の内容の差違がデータ表現に反映されず，異なる内容について記述しているWebページを同じカテゴリに分類してしまうことがある．Webページ間の差違を認識できる重要語は，負ノードと不正解ページに共通して出現し，正解ページには出現しない単語である．これらの重要語は，Webページ中で共起して出現している他の重要語との相互情報量において，最も高い値を持つ重要語の子ノードとしてベイジアンネットワークに付加される．

トポロジの変更

異なる問題領域について記述されているWebページが同一のカテゴリに分類される場合，両者で共通している属性値が誤分類の原因と考えられる．このような場合，ベイジアンネットワークにおいて，共通の属性値となっている重要語のアークを変更し，両者を分離しなければいけない．具体的には，次の手順に従って，ベイジアンネットワークのトポロジが更新される．

(1) 不正解ページと正解ページにおいて，ノイズとして除去されなかった重要語をそれぞれNword, Pwordに入れる．
(2) ベイジアンネットワーク中で，Nword中の重要語（Word）の親ノードにPword中の重要語が含まれていれば，WordをMoveに入れる．
(3) Moveに含まれている重要語（Mword）の親ノードを変更する．すなわち，正解ページにおいて，Mwordと共起して出現している他の重要語との相互情報量を計算し，最も高い値を持つ重要語をMwordの新たな親ノードとしてアークを付け変える．

共起出現確率の制御

重要語がWebページに出現していたとしても，重要でない内容を記述している場合には，その属性値を0とすることが望ましい．本システムでは，重要語の属性値が内容の重要度に応じて与えられるように，次の手順を用いてベイジアンネットワークを更新している．

(1) 不正解ページからノイズとして除去されなかった重要語をWordに入れる．
(2) Wordに含まれている重要語で，負ノード中のWebページに出現している重要語をUwordに，出現していない重要語をDwordに入れる．
(3) Uwordに含まれる単語の共起出現確率（$Prob_u$）と，Dwordに含まれる共起出現確率（$Prob_d$）を，それぞれ式(6)，式(6)に基づいて変更する．

$$NewP_u = Prob_u + 1/2(1 - Prob_u) \quad (5)$$

$$NewP_d = 1/2Prob_d \quad (6)$$

不正解ページと負ノードに共通している重要語は，不正解ページが負ノードに属するために必要な単語である．したがって，この語がノイズとして除去されることのないように，式(6)を用いて出現確率が大きくなるようにしている．一方，負ノードと共通していない重要語は，不正解ページが誤分類された原因となっている．このため，この語はノイズとして除去されるように，式(6)を用いて出現確率を低くしている．なお，式(6)の第2項の係数および式(6)の係数は確率を変動させる割合を意味しており，実験により導かれた数値である．

3. 本システムの実装

本システムは，文字処理言語Perlとリスト処理言語Prologを用いて実装している．ユーザが検索のキーワードとdesired textのURLを入力すると，システムは結果としてdesired textの内容と類似したWebページを提示する．提示されたWebページがdesired textと類似していない場合，不正解ページの*check box*をクリックすると，システムは「失敗からの知識獲得」を行う（**図5**参照）．具体的には，ユーザが不正解ページを指摘すると，個々のWebページから誤って検索された原因が特定され，ベイジアンネットワークが更新される．「失敗からの知識獲得」が終わると，最新のベイジアンネットワークを用いてWebページの再クラスタリングが行われる．上記の処理の後，システムは最初の検索画面に戻り，新たな検索を行うことができるようになっている．

4. 実験と評価

4.1 Webページの検索実験

提案したシステムの評価実験を行った．問題領域は，“Artificial Intelligence”とし，200個のWebページをデータとして用いている☆．また，初期ベイジアンネットワークは，実験で使用されたWebページを用いて構築しており，重要語としてTFIDFの上位200にランクされている単語を用いている．実験は，人工知能について研究している6名の学生に対して行った．なお，正解ページは被験者の研究分野に関連したWeb

☆ Webページの自動収集ロボットとして，検索エンジンAltaVistaを利用している．

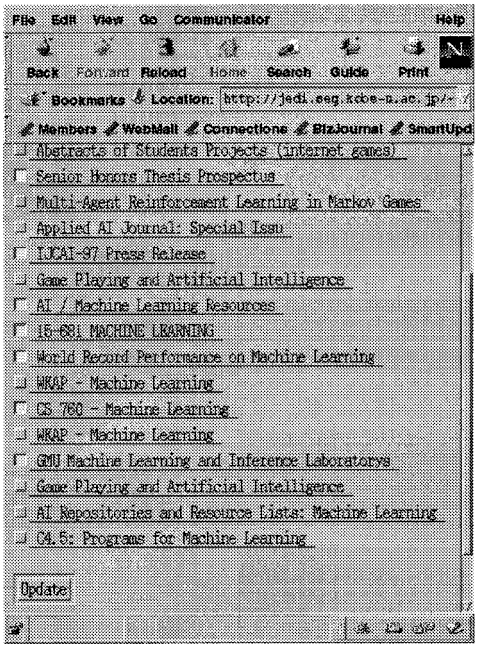


図 5 Web ページの検索結果
Fig. 5 Retrieved Web pages.

ページとし，desired text は正解ページの中からランダムに選択された Web ページとしている．実験では，desired text 以外の Web ページをクラスタリングして階層木を構築し，desired text の正解ページを抽出して検索精度を評価している．なお，評価尺度には再現率（Recall）と適合率（Precision）を用いている．また，検索結果に不正解ページが含まれていれば，ベイジアンネットワークを更新し，再クラスタリングをして新たな検索結果を被験者に提示している．

4.1.1 検索精度

被験者 6 名に対する実験の再現率と適合率の平均を**図 6** に示す．図 6 から，ベイジアンネットワークを更新する回数が増えるに従って，検索精度が向上していることが分かる．これは，漸増的なベイジアンネットワークの構築が検索精度の改善に有効であることを意味している．しかしながら，「失敗からの知識獲得」によって適合率は向上しているが，再現率はあまり改善されていない．これは，知識獲得の手法に原因があると考えられる．

本システムでは，不正解ページが desired text の属するカテゴリから排除されるためのみの知識獲得を行っている．逆に，desired text の属するカテゴリ以外に分類された正解ページが，次回からの検索で抽出されるために必要な知識獲得は行われていない．したがって，正解ページに対する知識獲得は積極的に行われないため，再現率があまり向上していないと考えられる．

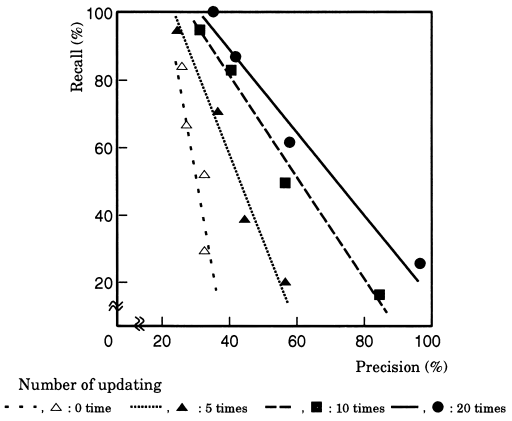


図 6 適合率と再現率
Fig. 6 Precision and recall.

4.1.2 データの順序が及ぼす影響

本システムでは，Web ページのクラスタリングとして逐次的な処理を行う COBWEB アルゴリズムを用いているため，構築される階層木はデータの順序に影響される可能性がある．そこで，6 回の実験では同じデータを使用し，各実験でのデータの順序は異なるようにした．この結果，ベイジアンネットワークの更新回数が少ない時点では，適合率に関して 14%の幅でばらつきが見られたが，更新回数が増えるに従ってばらつきは見られなくなった．このため，「失敗からの知識獲得」はデータの順序による検索精度への影響を軽減しているといえる．

4.1.3 シソーラスを用いた情報検索との比較

我々は，これまでに，検索の失敗に基づいた知識獲得による情報検索として，シソーラスを用いたシステムを提案してきた[6]．このシステムでは，シソーラスに基づいて Web ページをクラスタリングし，正解ページを検索している．このとき，不正解ページが検索されると，ユーザとの対話によってシソーラスを更新し，次回からの検索精度を改善するようにしている．従来のシステムの検索精度を**図 7** に示す．本実験で使用しているデータと，従来のシステムの評価実験で使用しているデータは異なっているため，直接比較することはできないが，図 6 と図 7 から，従来のシステムの方が検索精度が良いことが分かる．これは，従来のシステムが正解ページと不正解ページのいずれに関しても，シソーラスを更新しているためである．しかしながら，従来のシステムでは，ユーザが重要語の取捨選択や問題領域名などの入力を行う必要があり，負担

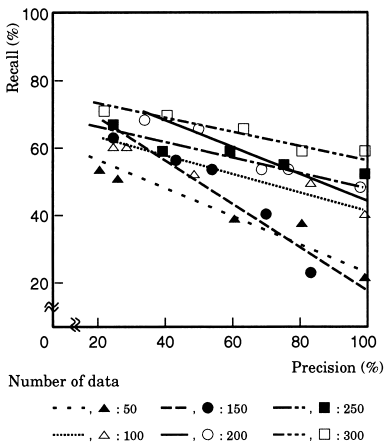


図 7 シソーラスを用いたシステムの検索精度
Fig. 7 Precision and recall of the previous system using a thesaurus.

の大きなものとなっている．また，詳しい専門知識を持たないユーザが使用することは困難である．本論文で提案したシステムでは，従来のシステムよりも検索精度が若干劣っているが，ユーザは提示された Web ページに興味を持つかどうかの評価のみを行うため，負担は少なくなっている．

4.1.4 構築されたベイジアンネットワーク

次に，実験で構成されたベイジアンネットワークを**図 8** に示す．なお，アークの幅は共起出現確率を表しており，太いアークは高い確率を意味している．

重要語の付加

図 8 から,「失敗からの知識獲得」により，いくつかの新たな重要語がベイジアンネットワークに付け加えられていることが分かる．たとえば，“email” の子ノードとして追加されている “Hillcrest”，“West”，“Frederick” は，会社の住所を表す単語である．これは，ソフトウェアの販売に関する Web ページに基づいて，ベイジアンネットワークが更新された結果である．この場合，製品で使用されているアルゴリズムが，ユーザの興味がある問題領域に関連していたため，製品販売の Web ページが正解ページとして検索された．しかしながら，実際には関心のあるアルゴリズムに関する詳しい記述はされておらず，不正解ページと評価されたため，異なるカテゴリに分類されるように，新たな重要語，すなわち “Hillcrest”，“West”，“Frederick” が追加されている．

トポロジの変更

また，トポロジの変更も 2 カ所行われている．たとえば，“information” の親ノードは “chess” から “retrieval” に変更されている．この場合，desired text が “information retrieval” に関する Web ページであるにもかかわらず，誤って “chess” に関するページが検索された．このため，ベイジアンネットワークが更新されて，“information” の親ノードが “retrieval” に変更されている．

共起出現確率の変更

さらに，重要語間の共起出現確率も多数変更されている．たとえば，“conference”，“workshop”，“press” の共起出現確率は低くなるように変更されている．これは，desired text として入力された興味のある論文に，その出典が記載されていたため，“Call for paper” が正解ページとして検索された．しかしながら，ユーザは “Call for paper” には関心がないため，不正解ページと評価されて共起出現頻度が変更されている．この更新により，ベイジアンネットワークの更新前では，“workshop” の出現確率が 0.92 と推論されていたが，更新後は 0.46 と推論されるようになった．本システムでは，重要語の属性値が 1 となる出現確率を 0.5 以上としているため，更新前の “workshop” の属性値は 1 となるが，更新後は 0 となる．また，更新前の “conference” の出現確率は 0.5，更新後は 0.25 と推論されるため，属性値は 1 から 0 に変更される．さらに，“press” の場合には，更新前の出現確率は 0.95，更新後は 0.42 と推論されるため，属性値は 1 から 0 に変更される．したがって，desired text と “Call for paper” では共通となる属性値がなくなり，両者が同じカテゴリに分類されることはなくなる．

4.2 ベイジアンネットワークの有効性

ノイズの除去

Web ページのノイズを除去する実験を行った．実験では，ノイズの除去を行わずに本システムを動作させた場合と，そうでない場合の検索精度を確かめている．実験方法および使用したデータは前節の実験と同じである．なお，**図 9** はベイジアンネットワークの更新回数が 20 回のときの検索精度を比較している．

図 9 から，ノイズの除去によって適合率が向上していることが分かる．適合率とは，関心のある Web ページが検索結果で占める割合を表す尺度である．したがって，実験の結果から，ノイズの除去により Web ページのデータ表現が内容に基づいたものとなり，クラスタリングが適切に行われていることが分かる．実際，ノイズの除去を行わずに Web ページを検索した場合には，リンク集などのような，問題領域に関する情報をあまり持たない Web ページが多く検索されていた．

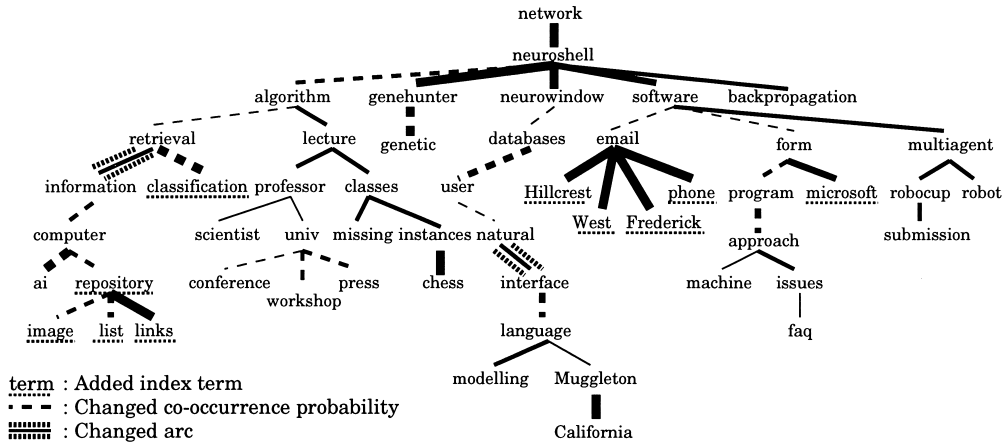


図 8 失敗からの知識獲得により構成されたベイジアンネットワーク
Fig. 8 Organized Bayesian network after the process of "knowledge acquisition by failure."

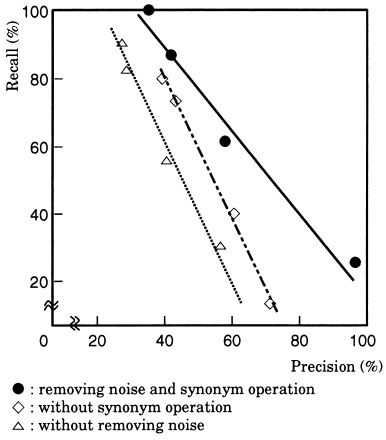


図 9 ベイジアンネットワークの有効性
Fig. 9 Efficiency test of a Bayesian network with synonymous information.

同義語の付加

次に，重要語の同義語を付加する実験を行った．実験では，同義語を考慮してデータ表現を行った場合と，そうでない場合の検索精度を確かめている．実験結果を図 9 に示す．

図 9 から，同義語の情報を付加することによって，再現率が向上していることが分かる．再現率とは，ユーザの希望する Web ページが検索された割合を示す尺度であり，この向上は，正解ページがより多く検索されていることを意味している．本システムでは，同義語を付加する際に，ベイジアンネットワークを用いて同義語と判断した重要語の属性値を 1 としている．このため，desired text と正解ページに共通となる属性値が増えたため，より多くの正解ページが desired text の属するカテゴリに分類されるようになったと考えられる．

4.3 失敗からの知識獲得の有効性

ベイジアンネットワークの更新に関する実験を行った．実験では，重要語の加加を行わずに本システムを動作させた場合，共起出現確率の変更を行わない場合，トポロジの変更を行わない場合，3 種類の更新を用いた場合の検索精度を確かめている．実験方法および使用したデータは前節の実験と同じである．なお，**図 10** はベイジアンネットワークの更新回数が 20 回のときの検索精度を比較している．

重要語の付加

図 10 から，重要語の付加によって再現率と適合率がともに改善されていることが分かる．実験の結果から，ベイジアンネットワークで認識できなかった話題に関する重要語を獲得することができている．このため，desired text の属するカテゴリから不正解ページが排除されるようになり，適合率が向上していると考えられる．また，付加された重要語の中には，正解ページに関する重要語も含まれていた．したがって，これまで検索されなかった正解ページも抽出されるようになり，再現率も向上したと考えられる．

トポロジの変更

一方，図 10 から，トポロジの変更によって適合率が向上していることが分かる．つまり，異なる問題領域に関する Web ページは desired text の属するカテゴリに分類されないようになったと考えられる．しかしながら，適合率の向上は数パーセントにすぎず，他の更新方法に比べると，検索精度の向上に対する貢献度は低い．これは，Web ページに記述されている内容

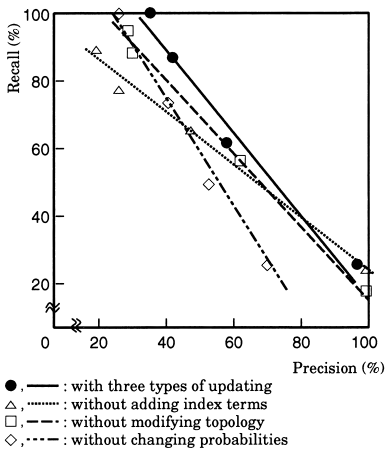


図 10 「失敗からの知識獲得」に関する有効性
Fig. 10 Efficiency test of updating a Bayesian network.

に応じて，最初からある程度正しく Web ページの分類が行われていることを意味している．つまり，属性となる重要語が，多くの問題領域からくまなく選択されているといえる．本システムでは，TFIDF の値が上位にランクされている単語を重要語として使用している．このとき，適切なしきい値を与えなければ，特定の問題領域に関連する重要語のみが選択される可能性があり，Web ページの誤分類が多発してしまうという問題がある．しかしながら，この場合には 2.6.3 項に従ってトポロジが変更され，最終的には Web ページが正しく分類されるようになると考えられる．

共起出現確率の変更

また，図 10 から，共起出現確率の変更によって適合率が改善されていることが分かる．つまり，Web ページが内容の重要度に応じて記述されるようになり，不正解ページが正しく分類されるようになったと考えられる．実際,「失敗からの知識獲得」において，ベイジアンネットワークの更新回数が少ない段階では，共起出現確率の変更は頻繁に行われているが，更新の回数に反比例して共起出現確率の変更回数が減少している．このため，重要語間の共起出現確率は興味のある問題領域の概念に沿うように収束していると考えられる．

5. 関 連 研 究

インターネットを介して Web ページを検索する手法として，さまざまな研究が行われている．たとえば，Syskill & Webert [9]ではユーザプロファイルを用いて Web ページを推薦するシステムである．ユーザプロファイルとは，ユーザの嗜好を記述したファイルである．システムは，検索した Web ページを提示して，ユーザにそれぞれ「興味がある」か「興味がない」を判定してもらう．これらのページを訓練例として，全ページに含まれる語のうち information gain の大きなものから順に選択して，ユーザプロファイルの学習を行っている．Syskill & Webert では，ユーザプロファイルが Web ページの判断に大きく依存しているにもかかわらず，同義語やノイズのことを考慮していないという問題がある．

テキスト文書を解析するために，Xu [11]らは名詞グループに対する重みを導入している．名詞グループとは，ユーザからの質問文中の共起に基づいて選択された単語の集合であり，重みは検索された文書の中で上位にランク付けされている文書を用いて修正される．このため，システムの挙動は上位にランクされた文書の影響を受けるという問題がある．具体的には，質問文に従って検索された文書が互いに意味的に類似していない場合には，同義語を考慮していないためにシステムがうまく機能しないという問題がある．

ネットワークニュースをフィルタリングするために，ニューラルネットワークを用いたシステムも提案されている[8]．このシステムでは，ニューラルネットワークのノードにニュース中の単語が割り当てられ，ノードのエネルギーやノード間のリンクの重みによって，ユーザの興味が学習されるようになっている．たとえば，ニュース中の単語が共起している場合は，対応したノード間のリンクの重みが増やされるようになっている．

6. お わ り に

本論文では，Web ページの検索の失敗と個人の嗜好に着目し，ベイジアンネットワークを用いてクラスタリングを行う情報検索を提案した．実験により，検索の失敗に基づいてベイジアンネットワークを更新すれば，検索精度の向上に有効であることを示した．

本システムでは，領域知識を表現するためにベイジアンネットワークを導入している．このため，シソーラスを使用せずに同義語の処理ができるようになっている．しかしながら，ベイジアンネットワークで兄弟関係となるすべての重要語が，同義語の関係になるとは限らない．なかには，意味的な類似性を持たない重要語が兄弟関係になることもある．これは，ベイジアンネットワークが自動的に構築され，漸増的に更新されていくことが原因であると考えられる．同義語を特定する手段としてシソーラスを使用する方法もあるが，データの準備や保守といった負担がある．このため，利用する際の負担がなく，同義語の特定が可能なベイ

ジアンネットワークを導入することは有効であると考えられる．

本システムでは，Web ページの自動収集ロボットとして AltaVista を使用している．一般に，検索エンジンの検索結果は膨大な数となってしまうため，単独で検索エンジンを用いても正解ページを発見することは困難である．実験では，AltaVista によって検索された Web ページをさらに本システムを用いて取捨選択し，正解ページに関する検索精度の向上を達成している．このように，本システムと既存の検索エンジンを組み合わせて多段階の情報検索を行えば，正解ページを効率良く抽出することができると考えられる．

さらに，desired text と意味的な関連が少ない正解ページを検索することは困難である．このような正解ページを検索するために，ユーザの嗜好を記述しておくユーザプロファイルを利用する手法がある[2]．ユーザプロファイルは，検索された文書に出現している単語を用いて記述され，ユーザによる修正が行われる．しかしながら，このような操作はユーザの負担を大きくするという新たな問題が生じてしまう．このため，本システムでは desired text と不正解ページを分離するために必要な知識獲得のみを行っている．

本システムでは，正解ページを検索する際の評価として，ベクトルのユークリッド距離のみを利用している．Freitag[5]は，多戦略による複数の評価を利用して情報検索の精度を向上させている．このように，頑強な自然言語処理や述語論理を用いた推論ルールを用いれば，より多くの正解ページを検索できるようになると考えられる．

本研究では，初期ベイジアンネットワークの構築および更新，ノイズの除去，Web ページの検索過程などでしきい値を設定している．すべてのしきい値は実験を繰り返して経験的に得られた値である．これらのしきい値の妥当性は実験によって経験的に得られたのみであるが，データによっては大きく変わるものもあると考えられる．今後は，多くのデータに対して本システムを適用し，自動的にチューニング可能なメカニズムを開発する必要がある．

最後に，本実験では検索の精度を評価するために再現率と適合率を用いている．しかしながら，ある Web ページが正解ページかどうかはユーザによって決定されるため，検索領域についてあまり詳しい知識を持たないユーザが本システムを利用した場合，再現率と適合率の評価尺度が有効に働かないことがある．今後は，情報検索を正しく評価するための評価尺度についても研究する必要がある．

謝辞 本研究は，一部，文部省特定領域研究「発見科学」の援助を受けている．

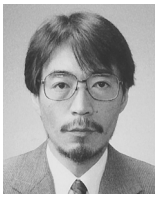
参 考 文 献

1) Balabanovic, M. and Shoham, Y.: Learning Information Retrieval Agents: Experiments with Automated Web Browsing, *Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous*, pp.13–18 (1995).
2) Bloedorn, E., Mani, I. and MacMillan, T.R.: Machine Learning of User Profiles – Representational Issues, *Proc. 13th National Conf. on Artificial Intelligence*, pp.433–438 (1996).
3) Bookstein, A., Klein, S.T. and Raita, T.: Clumping Properties of Content-Bearing Words, *Journal of American Society for Information Science*, Vol.49, No.2, pp.102–114 (1998).
4) Fisher, D.: Knowledge Acquisition via Incremental Conceptual Clustering, *Machine Learning*, Vol.2, pp.139–172 (1987).
5) Freitag, D.: Multistrategy Learning for Information Extraction, *Proc. 15th Int. Conf. on Machine Learning*, pp.161–169 (1998).
6) Horii, N. and Uehara, K.: Incremental Clustering for Dynamic Information Filtering, *Proc. Int. Symposium on Digital Media Information Base*, pp.79–86 (1997).
7) 伊藤哲郎：情報検索，昭晃堂 (1986).
8) Jennings, A. and Higuchi, H.: A Personal News Service Based on a User Model Newral Network, *IEICE Trans. Information and Systems*, E75-D (2), pp.198–209 (1992).
9) Pazzani, M., Muramatsu, J. and Dillsus, D.: Syskill & Webert – Identifying Interesting Web Sites, *Proc. 13th National Conf. on Artificial Intelligence*, pp.54–61 (1996).
10) Shoham, Y.: *Artificial Intelligence Techniques in Prolog*, pp.165–197, Morgan Kaufmann (1994).
11) Xu, J. and Croft, W.B.: Query Expansion Using Local and Global Document Analysis, *Proc. 19th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp.4–11 (1996).

(平成 11 年 3 月 8 日受付)
(平成 11 年 11 月 4 日採録)

堀井 則彰（正会員）

1975 年生．1997 年神戸大学工学部情報知能工学科卒業．1999 年同大学院自然科学研究科情報知能工学専攻博士前期課程修了．同年松下電器産業（株）入社．現在，ディジタル音響処理の研究に従事．



上原 邦昭（正会員）

1954 年生．1978 年大阪大学基礎工学部情報工学科卒業．1983 年同大学院博士後期課程単位取得退学．大阪大学産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授を経て，同大学都市安全研究センター教授．情報知能工学科を兼任．工学博士．人工知能，特に機械学習，マルチメディアデータベース，自然言語によるヒューマンインタフェースの研究に従事．1990 年度人工知能学会研究奨励賞授賞．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．