

ユーザの変化する興味を理解し表現する文献検索支援システム Index Navigator

An Index Navigator: Understanding and Expressing User's Changing Interest

大澤 幸生* 須川 敦史* 谷内田 正彦*
Yukio Ohsawa Atsushi Sugawa Masahiko Yachida

* 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531 Japan.

1997年5月21日 受理

Keywords: information retrieval, user's changing interest, abduction, distributed reasoning.

Summary

It is important to guide a user to interesting documents in a large-scale document-database. However, when the user is not an expert of the area of his/her new interest, it is difficult for the user to name precise keywords in which he/she is interested, nor to select areas of his/her own interest. This paper presents an Index Navigator which clarifies what areas the user is interested in, what keywords he/she should investigate, and what documents concern his/her interest. A tough problem for such a system is to understand interesting areas for the user, among other areas-sets which can explain his/her behaviors. Our Index Navigator employs an inference method called Cost-based Cooperation of Multiple Abducers (CCMA), for understanding user's interest from the history of the user's expression of interest in insufficient keywords, even if the changing speed of the user's interest is totally unknown. With this device, the Index Navigator guides the user to really important areas, keywords and documents.

1. はじめに

市販のCD-ROMから電子図書館・インターネットに至るまで、近年の電子文書の増加は著しい。これらの文書データベース（以下、文書DBと略）からユーザが欲しい文書を得るため、各種のサーチエンジンが作られてきた。ユーザの興味をキーワード群で表し、文書DB中の文書のキーワードベクトル（どの単語がどれくらい頻りに現れるかという情報）と照合するのがサーチエンジンの主流であり、その発展型としてユーザが読み書きした文書のキーワードベクトルと文書DB中の文書のキーワードベクトルを照合するアプローチもある [Krulwich 95]。

しかし、ユーザと文書DB中の文書が同じ意味を異なる語で表現するような場合は、この照合が難しくなる [Furnas 87]。特に、ユーザが何か新しい知識を得

るために参考文献を探すような場合は、目標の文書がユーザの知らない言葉で書かれていることが多いので、ユーザと文書DBの語彙の相違が著しくなる。ユーザの心理的な都合（いわゆる度忘れなど）のために表現力が低下している場合も、同様の問題が生じる。

そこで筆者らは、ユーザが自分の興味をキーワードで表す作業を支援する、文献検索支援システム **Index Navigator** を研究・開発した。Index Navigatorの目的は、ユーザが指定した不十分な（すなわち、文書DB側の検索用キーワードをユーザの興味を明らかにできるほど多く含まない）キーワード群からでもユーザの興味を理解し、これによりユーザの興味に関連の深いキーワードをユーザ自身に教えることである。ユーザは、システムから自分に関係の深いキーワードを教わりつつ文献検索を行えるので、スムーズに文献検索を行い、該当分野の重要な概念に親しめるようになる。

特に本論文では、ユーザが、興味は持っているが専

門外の分野の文書 DB から自分の興味に合う文献を探すのを支援する Index Navigator を考える。この場合、ユーザが当該分野の用語を知らないことが多いので、キーワード照合における上記の問題は深刻となる。

一つの難しい問題は、ユーザの興味をいかにして推論するかである。ユーザの指定できるキーワード群が不十分では、ユーザの興味を十分絞り込めないことになる。この不足を補う一つの方法は、ユーザがある瞬間に指定したキーワード群だけでなく、ユーザが指定したキーワード群の履歴も考慮して推論を行うことである。各時点で指定されたキーワードの不足を、ユーザの現在と同じ興味を持続していた過去の時点で指定されたキーワードによって補うのである。

では、ユーザの興味を持続している期間はどのようにすればわかるであろうか。ユーザの興味を理解する手法として、ユーザの過去の興味から学習したユーザモデルを用いて未来の興味を予測するアプローチがある [Armstrong 95, Krulwich 95]。しかしユーザの興味の変化まで考えると、計算機とユーザ、あるいはその相互作用のいずれでもない外界からの想定外の影響（例えば、人との会話や紙上の文章）による興味の急変が実際には頻繁なため、ユーザの過去の履歴からは現在・未来の興味を推察できないことが多い。

ユーザモデルを時刻と共に変動させる方法もあるが、ユーザの興味の変化する時点を検出する手法を確立しなければ、いつユーザモデルを変化させるかという問題が残ってしまう。この理由で、例えばユーザによる各時点でのさまざまな単語の使用頻度を一時的なユーザモデルとする手法 [Balabanovic 95] は、キーワードの不足という先の問題を解決できない。なぜなら、例えば日本語から英語への翻訳に興味を持つユーザが、「翻訳」という単語の使用を減らして「英語」や「日本語」の使用頻度を上げたからといって「日本語から英語への翻訳」への興味は下がっておらず、むしろ「翻訳」、「英語」、「日本語」の3単語が、同じ興味の持続している期間に指定されたことが重要だからである。

Index Navigator では、ユーザの興味を仮説推論のコストに基づく協調 (Cost based Cooperation of Multiple Abducers, 略して **CCMA**[大澤 97b]) により理解する。CCMA は変化の速度が未知な時系列を理解するための推論であるので、ユーザの興味の予想外の変化を理解するのに適している。これにより Index Navigator は、各時点でのユーザの興味の対象分野を理解し*1、その結果、ユーザの興味に関連の深いキーワー

*1 ユーザの興味の分類自体も、図書館で文献調査を行ったリ、参加する学会を選択するのに役立つであろう。

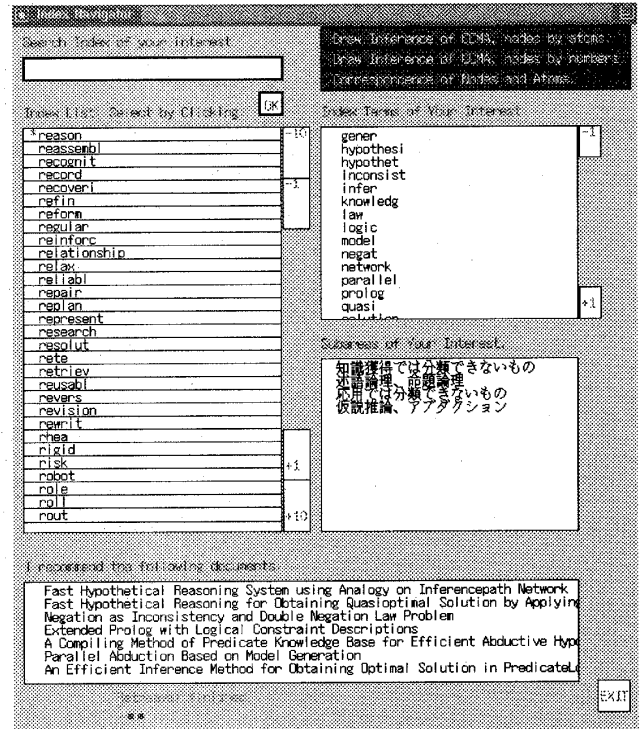


図1 Index Navigator のウィンドウ

ド群と文書を出力することができる。

2. Index Navigator の概観

Index Navigator のユーザは、索引 (図1左側の大きなウィンドウ) 上のクリックか、左上の小さなウィンドウへのキーワードの書き込みによって、興味を惹かれるキーワードを選択する*2。各時点でユーザが指定したキーワードの履歴から、ユーザの興味ある分野を推論するのが Index Navigator における先述 (詳しくは後述) の CCMA の役割である。ここで推論された分野が分野リスト (図1右中段) に書き込まれ、これらの分野からユーザの興味に関連の深いキーワード群が導かれキーワードリスト (図1右上段) に出力される。分野リストからは同時に、ユーザの興味に合う論文も導かれ文書リスト (図1最下段) に出力される。

3. Index Navigator の動作原理

この章では、Index Navigator の次の2つのサブシステムの動作原理を詳説する。

2 図1の索引を上下にスクロールすると他のキーワードを表示することができ、ユーザがいくつかの単語に一時点で興味を持っている場合は同時に複数のキーワードが指定可能である。図1の画面では、“”のついた“reason”が指定されたキーワードの一つである。

- 1) ユーザの興味理解部：ユーザが指定したキーワード群の履歴から、ユーザの興味の対象分野を推論する。
- 2) キーワード・文書呈示部：1) で求めた分野の中から、ユーザの興味に関連の深いキーワード・文書を求めて呈示する。

3.1 ユーザの興味理解部

ユーザの興味は、外界からの影響により予想外の変化を遂げて行く。そこで、Index Navigatorでは、ユーザの興味の変遷の様子を CCMA[大澤 97b] によって理解する。CCMA は、変化の速さを（全く不明なために）知識中に与えることができないような時系列の原因を推論する枠組みである。

[1] CCMA の原理

CCMA の機能は、事象の時系列の一貫した原因と、その一貫性が破れる時点、すなわち原因の時間軸上の変わり目を見つけることである。

時系列の一貫性を捉える推論は、CCMA 以前にも研究されてきた。例えば、マルコフモデルにおける状態遷移確率は、時刻の変化につれて状態がどの程度の速さで変化するかを確率値によって示したものとなっている。また、ベイジアンネットワークにおいては、事象間の因果関係の強さが条件付き確率として与えられており、これを時点間の因果関係に適用すると一貫性をモデル化したことになる [Kashino 95]*³。しかし [大澤 97b] に指摘したように、このような時系列の一貫性の確率モデルは、変化の速度が全く分からない時系列事象の原因は推論できない。

一方、CCMA[大澤 97b] は、時点間の遷移確率のような一貫性が見積りが与えられない場合に、複数のアブデューサ（後述の、コストに基づく仮説推論を行うシステム）のそれぞれに一時点の観測事象の原因を推論させ、隣接する時点のアブデューサ間の協調によって時間方向に一貫性のある原因を推論するものである。この協調推論により、原因の一貫性と原因の切れ目を検出する様子を次に概説する。

● CCMA の推論プロセスの概要

連続する N 時点 (t_1, t_2, \dots, t_N) からなる時系列中の各時点での観測事象を $G_j (j = 1, 2, \dots, N)$ で表す。 $j = 1, 2, \dots, N$ の各観測事象 G_j の原因を、時間方向に一貫性を持つ仮説 $h_j (j = 1, 2, \dots, N)$ として得るのが

*3 情報検索の場合、ベイジアンネットワークの条件付き確率をユーザの興味の時間方向の遷移確率として用いたものよりも、ある時点での分野におけるキーワード出現頻度として用いたものの方が有意な成果を多く挙げている [Fung 95]。

CCMA の役割である。ただし、 h_j は仮説集合 H の部分集合で (式 (1))、背景知識 Σ と合わせて G_j を導くことができ (式 (2))、 Σ と矛盾しない (式 (3)) 解仮説でなければならない (ϕ は空節)。

$$h_j \subset H, \quad (1)$$

$$h_j \cup \Sigma \vdash G_j, \quad (2)$$

$$h_j \cup \Sigma \not\vdash \phi. \quad (3)$$

Σ は、式 (4) のようなホーン節を集めた背景知識である。 X_1, X_2, \dots, X_n は条件、 Y は結論を意味するアトム（知識の記述単位）で、本論文ではそれぞれ文献 DB 中の分野と、文献 DB で分野 X_1, X_2, \dots, X_n の全てに含まれるキーワードを表す。

$$Y :- X_1, X_2, \dots, X_n. \quad (4)$$

Y が空節 ϕ ならば式 (4) は条件のアトムが互いに矛盾することを意味し、この矛盾は式 (3) で禁止されている。

アルゴリズムの詳細は [大澤 97b] に譲るが、CCMA では、時系列事象の一貫性のある原因を以下の協調推論によって理解する。まず、各時点のゴール（観測事象）にそれを説明するアブデューサを割り当てる。ここで j 番目のアブデューサは、式 (1)~(3) の制約を満たしコストの最小な解仮説 h_j を得る推論（コストに基づく仮説推論：Cost based Abduction, 以下 CBA と略す）を行なう。ここで解仮説 h_j のコストとは式 (5) の $cost$ である。 $w(\eta)$ は H 中の仮説 η の重みで、 η が真となる確率 $p(\eta)$ に対し $w(\eta)$ を式 (6) のように与えると、CBA の解仮説 h_j は G_j の事後確率最大の説明となる [Charniak 94]。

$$cost = \sum_{\eta \in h_j} w(\eta). \quad (5)$$

$$w(\eta) = -\log p(\eta). \quad (6)$$

CCMA では、時点 t_j 番目のアブデューサが自分の解仮説 h_j を両隣のアブデューサにメッセージとして伝える。このメッセージの伝え方が CCMA の特徴で、解仮説 h_j 中の各仮説と同じ仮説の隣接（時点 t_{j-1} と t_{j+1} の）アブデューサにおける重みを 0 とするのである。これは式 (6) から、 t_j のアブデューサの解仮説が真であることを隣接時点でも強く信じる意味を持つ。そして、各アブデューサは再び CBA の推論を行ない、解仮説を得て再び両隣にメッセージを出す。この繰り返しを、どのアブデューサのどの仮説の重みも変化しなくなるまで繰り返す。

各仮説の重みを初めに与えるから一時点で仮説が真となる確率は用いることになるが、時点間の状態遷移確率は不要である。従って、CCMA は状態の変化の速

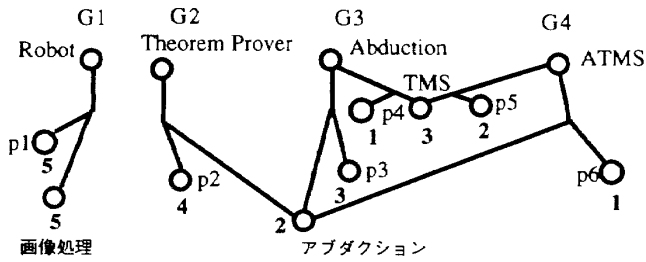


図2 ユーザの興味理解部でCCMAを適用する知識の例

さが未知の時系列に適用可能となる。また、状態(どの仮説が真であるかという組合せ)が変わる時点では、ゴールを説明できない仮説は前後の時点の解仮説の影響で重みを0にされても解仮説から却下されるので、多くの場合この変わり目を正しく得ることができる。

CCMAと同様に協調推論の枠組みから事象の時系列を理解する研究には他にFA/C[Carver 95, Carver 96]などがあるが、CCMAとは異なり、全ての時点の観測事象から時点間の矛盾がない(すなわち、変化しない)原因を仮説として求めるものである。

● ユーザの興味理解部におけるCCMAの推論例

図2最上段の(根の)ノードが各時点のゴール(指定されたキーワード)を表し、最下段の(葉の)ノードは仮説(ユーザの興味の対象かも知れない分野)を表す。例えば、図2のG1:“Robot”,G2:“Theorem Prover”,...はそれぞれ時刻 $t_1, t_2 (> t_1), \dots$ でユーザの指定したキーワードであり、「画像処理」はユーザが画像処理の分野に興味を持っているという仮説を表している。この「画像処理」の重み(図2のノード「画像処理」の横の太字“5”)は、式(6)より文書DBにおける分野「画像処理」の文書の割合の負の対数とする。

また、図2中の枝は「キーワード ← 分野」という因果関係の規則をホーン節で表す。例えば、図2左端のグラフはホーン節

$$Robot :- \text{画像処理}, p1 \quad (7)$$

を表し、 p_1 は「画像処理分野の文書は $p(p_1)$ の確率で“Robot”を含む」という条件付き確率 $p(p_1)$ を表す形式上の仮説で、 p_1 の重みは式(6)より $-\log p(p_1)$ である。

このようなホーン節の集合で表された知識を元に、CCMAのそれぞれのアブデューサは、ユーザがそれぞれの時点でキーワードをどういった分野への興味から指定したのかをCBAによって推論する。そして、アブデューサ同士で先述のように仮説の重みを操作する協調によって、ユーザがどの期間でどの分野への興味を持続させたかという理解に達する。

図2のようにゴールの時系列と知識が与えられた場合

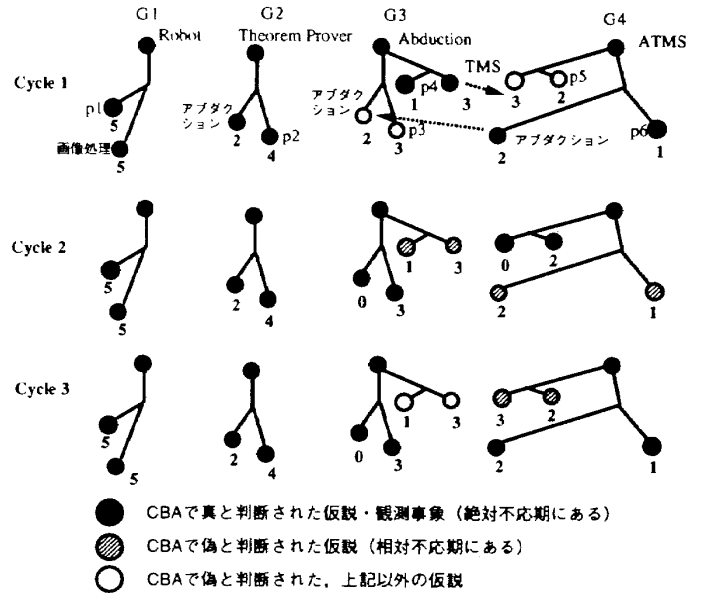


図3 図2の知識と時系列に対するCCMAの推論プロセス。不応期については[大澤 97b]参照

に、CCMAで行なわれる推論プロセスを図3に示す。まずCycle 1で、アブデューサらは解仮説としてユーザの興味が $\{h_1: \text{画像処理}, h_2: \text{アブダクション}, h_3: \text{TMS}^*4, h_4: \text{アブダクション}\}$ (h_i が、時刻 t_i における興味の分野を示す)と変化すると推論する。すると、メッセージが図3破線矢印のように伝達される。すなわち、Cycle 1のG4に対する解仮説 h_4 中の仮説に分野「アブダクション」があるので、これと同じ仮説「アブダクション」の重みを隣接(時点 t_3 の)アブデューサにおいて0にする。言い換えると、時点 t_3 のアブデューサは、 t_4 のアブデューサからのメッセージを受けて「ユーザはアブダクションに興味を持っているに違いない」と予想してCycle 2でCBAを再行する。この結果、Cycle 2の推論によって t_3 のアブデューサは、Cycle 1で解としていた「TMS」を含むコスト4の解仮説から、よりコストの低い(コストがCycle 1での5からCycle 2で3に減った)「アブダクション」を含む解仮説に更新する。その後Cycle 3で、どの仮説の重みも変化しなくなり推論が終了する。

以上の結果、このユーザの興味が一度「画像処理」から「アブダクション」に変化した時点 t_2 の後には興味を持続したと理解される。こうして、興味の一貫性と変化が検出される。

なお、詳細は[大澤 97b]に譲るが、図3の絶対不応期と相対不応期は、仮説の重みが隣接アブデューサか

*4 TMSは真理維持システム(Truth Maintenance System)の略で、ATMSなどのアブダクションの推論システムを含む。

らの影響を受ける度合いが弱められる期間であり、両不応期の併用によって協調推論が無限ループに陥るのを防ぐために導入されたものである。

以上の CCMA の推論によって、各時点（現時点を含む）でユーザの興味の対象となっていたであろう分野の集合がそれぞれ解仮説として求められ、現時点の興味として推論された分野が分野リストに出力される。

3・2 キーワード・文書の呈示部

次に分野リスト中の分野の集合から、キーワードリストと文書リストに出力するキーワードと文書が、それぞれ演繹的に求められる（式(7)の「画像処理」から“Robot”を得るように）。これらのキーワードと文書は以下の優先度によってソートされ、優先度の高いものから順にそれぞれキーワードリストと文書リストの上から順に出力される（図1参照）。

式(8)で、 p_0 の確率 $p(p_0)$ は $area_1, area_2$ の両分野に属する文書のうちキーワード $keyword_0$ を含むものの割合であり、 p_0 には重みとして $-\log p(p_0)$ が与えられる。すると、 $area_1$ と $area_2$ が分野リスト中にある場合には $keyword_0$ に優先度 $p(p_0)$ が与えられる。また、 $area_1$ しか分野リスト中になければ $keyword_0$ の優先度は $p(p_0) \cdot p(area_2)$ となる（ $p(area_2)$ は、文書 DB 中で $area_2$ 中の文書の割合）。したがって、分野リスト中の分野の多くがホーン節の条件と一致するほど結論部のキーワードの優先度が高くなる。

文書にも同様に、確率を基に優先度が与えられる。このために式(9)の p_1 の確率 $p(p_1)$ を、 $document_1$ が文書 DB における $area_1, area_2$ の両方に属する総数 M の文書中の一文書である（さもなくば式(9)は背景知識 Σ に含まれていない）ことから $1/M$ で与え、キーワードと同様に $document_1$ の優先度を計算する。

$$keyword_0 : -area_1, area_2, p_0. \quad (8)$$

$$document_1 : -area_1, area_2, p_1. \quad (9)$$

4. Index Navigator の評価実験

4・1 実験の方法

Index Navigator のシステムは C++ で構築した。実験環境としては、各被験者の作業場所にあるメモリ 64～128MB、CPU 133～200MHz の計算機を用いて操作した。以下が実験内容である。

● **文書データ** 人工知能学会誌十周年記念 CD-ROM（人工知能学会誌、Vol.11, No.3 の付録）中の論文および解説論文計 741 編。各論文がどの分野に属するかと

いう情報を含めて HTML 形式で蓄積されている。

ここから Perl5.0 で文書を取りだし、自動的に式(7)のようなホーン節の集合を生成することにより知識ベースを構築した。この知識ベースの規模は、 H 中の仮説数（分野数）が 95、 Σ 中のホーン節数が 6212、キーワード数が 2672 となった。

索引中のキーワードは、意味の近い派生語をまとめて扱うため、タイトルから [Swaminathan 93] と同様に助詞や冠詞を取り除き [Porter 80] の手法で語尾変化を除いたものとした。

● **ユーザ（被験者）** 人工知能が専門外の学生 10 名。
● **実験方法** 各被験者に図 1 と同様のウィンドウを同時に 4 つ見せ、それぞれユーザの興味理解部に **A) CCMA, B) 学習機能***⁵ 付きベイジアンネットワーク、**C) CBA**（各時点のユーザの興味を全く独立に CBA で推論させる）、**D) Explanatory Coherence**（[大澤 97b] で CCMA に対する従来手法として比較した [Ng 90] の仮説推論で、時点間に強い一貫性を仮定している。以下、**EC** と略）を用いた Index Navigator のインタフェースとした。**A)～D)** はウィンドウの上部バーの見出し（図1では“Index Navigator”）で区別したが、4通りの推論手法の名前は伏せてある。

各被験者は、各時点で 4 つのウィンドウのいずれかからキーワードを指定し、その結果得られたキーワードを用いて他のデータベースから得た論文を読んでも、Index Navigator が出力した論文を読んでも良い。そして、新しく興味を持ったキーワードを入力し、同じことを繰り返していく。ユーザが入力したキーワードを、各時点でのユーザの興味として自動記録した。

被験者が多くの文献を求めながら作業に集中するように、興味を持った分野の内容のレポートを実験期間（1週間とした）後に提出してもらった。

4・2 数値計測による Index Navigator の結果

各ユーザについて、各推論手法の表示内容と入力（興味を持ったキーワード）との一致の度合いを各入力時点で計測した。具体的には、図 4 と図 5 はそれぞれ、式(10)の Precision P と式(11)の Recall R を計測したもので、各時点のドットは、各推論手法についてユーザ全員の前後 3 時間の P, R を平均した結果である。式(10),(11)で、 $output$ は Index Navigator がユーザの興味として表示した分野およびキーワードの集合、 $input$ はユーザが直後に入力したキーワード⁶⁾の集合を表

*5 ユーザの興味の分野間の遷移確率を、実験開始時からの履歴から自動計算した。

*6 分野名もキーワードとして指定でき、その場合はその分野は自動的に解仮説に加えられる。

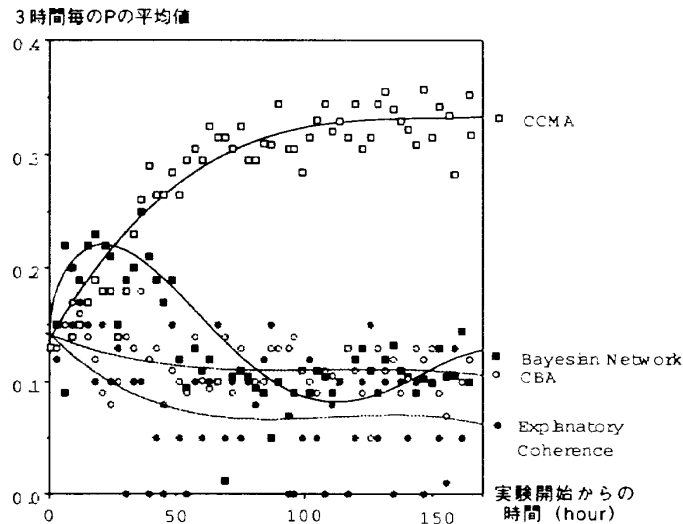


図4 各推論手法による解の精度 (Precision)

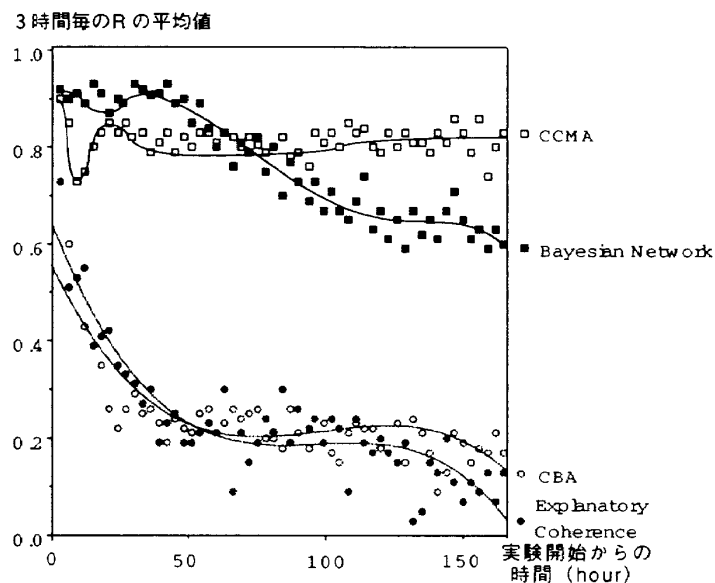


図5 各推論手法による解の精度 (Recall)

し、 $|Set|$ は集合 Set の要素の個数を表す。

$$P = \frac{|output \cap input|}{|output|}, \quad (10)$$

$$R = \frac{|output \cap input|}{|input|}. \quad (11)$$

図4から分かるように、CCMAを用いたIndex Navigatorでは、システムの指定したキーワードのうち直後にユーザの指定するものの割合が、時間と共に増加する。これは、ユーザがキーワードの意味を学習するにつれて、CCMAを用いたウィンドウA(4.1節参照)の出力キーワードのうち、自分の興味を表していると感じるものが増加したために、それらのキーワードを次に指定することが多くなったと解釈できる。

また、図5から分かるように、CCMAを用いたウイ

ンドウAは、ユーザの興味の多くの部分を安定して表示する結果となった。これは、ユーザの指定するキーワードのうち、ウィンドウB,C,DよりもウィンドウAの出力から選ぶ部分が多くなったのが主な原因と考えられる。しかし、全推論手法でのRの平均値は低下傾向にあった。これは、ユーザがキーワードを学習したため、Index Navigatorの出力だけに頼る必要が弱くなったためであろう。それでもCCMAの結果が比較的安定して高い値を示すのは、ユーザが自分の学習成果以外に頼るとすれば、他の推論手法による結果よりはCCMAの結果に好んで頼ったためと見られる。

このようにCCMAは、長い時間にわたってユーザの興味を適切なキーワード集合で表すのに適した推論であることがわかる。実験開始時直後はベイジアンネッ

トワークの性能が上回っているが、これは遷移確率の学習が、初めのうちはユーザの興味の実際の変化と一致したからと見られる。

一方、[大澤 97b]において示したことであるが、CBAは時点間の一貫性を考慮していないし、ECは強すぎる一貫性を仮定している。このため、CBAとECをそれぞれ用いたIndex NavigatorのウィンドウC,Dからキーワードを指定するユーザは、2日目までになくなった。すなわち、CBAではユーザの指定したキーワードの履歴を考慮しないので、筆者らが取り上げた問題である「単一の時点で指定されるキーワードの情報不足」について何ら解決していない。したがって、不足した情報からユーザの興味を推論することになるため、推論の誤りが多くなりユーザから「Index Navigatorは自分の興味を理解できない」と評価されてしまう。逆にECでは、ユーザの興味が変わることを考慮していないため、既に失われたユーザの興味まで各リストに出力される。その結果、ユーザにすれば「自分の興味の変化をIndex Navigatorが理解しない」ということになる。以上のCBAとECに対する評価は、実際のユーザの声とも一致した。

4.3 出力例による各推論手法の比較

比較的良好な経過を示したCCMAとベイジアンネットワークを比較すると、ベイジアンネットワークの方がユーザの興味の変化の解釈を誤ることが多い。例えば、あるユーザが2日目のほとんどの時点で分野「アブダクション」に興味を注いだ場合、3日目にこのユーザがキーワード“ATMS”を“abduct”^{*7}の次に選んでも分野「アブダクション」への興味だけが持続していると解釈された。しかし、実際はこのユーザの関心は3日目の途中からアブダクションの高速アルゴリズムへと変化していた。この例は、2日目までに学習した興味の変移確率を3日目も用いることが過剰な先入観となったことを意味している。一方CCMAでは、3日目のキーワード“ATMS”の指定以降のユーザの興味に分野「探索アルゴリズム」を含めることができた。

一方で、変化を鋭く捉え過ぎると、ユーザの興味は持続しているのに変化したと誤って解釈する危険がある。しかし、この誤りはCCMA使用時の感想としては報告されなかった。実際CCMAでは、“reason”を指定した(図1の出力になる)直後に“abduct”を選択すると「アブダクション」への興味は持続している

と理解したし、“reason”の次に“model”を選択すると、新しい「帰納学習」への興味と同時に、持続している「アブダクション」への興味も把握できた。

4.4 CCMAの限界

3.1節に示したように、CCMAでは、1サイクルでは隣接アブデューサ間でのみメッセージを送り合う。したがって、一貫した興味が続く時区間 $[t_0, t_3]$ 中の時区間 $[t_1, t_2]$ ($t_0 < t_1 \leq t_2 < t_3$)で前後と全く異なる興味が入り込められることがあると、 $[t_1, t_2]$ でのアブデューサの解仮説が前(t_0)と後(t_3)のアブデューサの仮説と異なるために、 t_0, t_3 のアブデューサの間のメッセージのやりとりを阻むことになる。これを、各アブデューサから $L(> 1)$ 時点前後のアブデューサまでメッセージを送るようにしても、 $[t_1, t_2]$ の長さは L より長いかも知れないから本質的な解決にはならないし、 t_3 以降のユーザの実際の興味が t_0 以前と違う場合には正しい推論を阻害することになる。

しかし、我々が扱っている、分野をある程度限定したデータベースからの検索(1章参照)では、前後と全く異なる興味が一貫した興味が続く中で挿入されることは少ない(インターネット全体を検索するような場合は別である)。というのは、全く違う興味を満たすには別のデータベースを用いることが多いからである。

そこで、何らかの重要な興味に関する仮説 η が t_0 と時区間 $[t_1, t_2]$ で共通しているとすると、 η を真とする t_0 のアブデューサのメッセージは時区間 $[t_1, t_2]$ を経て t_0 から t_3 まで伝わり、 t_3 での解は η を含むものにしぼり込まれる。すなわち、 t_3 でのアブデューサは「 t_3 で指定されたキーワードを説明できる」と「 η を含む」の両方でしぼり込まれた仮説を得るから、 t_0 と t_3 で同じ興味からキーワードが示されるなら、 t_0 と t_3 の解仮説が同じになる可能性は高くなる。

ただし、時区間 $[t_1, t_2]$ での興味が直前・直後と全く異なることもまれには起きる。その場合、[大澤 97b]における回路故障診断の結果にも見られたように、一般には状態の変わり目の直前・直後の解仮説の精度が劣化する。しかし、[大澤 97b]と同様に、興味の変り目から2~3時点以上一貫した興味が続けば、その時点からCCMAの解の精度は向上した。

このCCMAと比較すると、例えばベイジアンネットワークでは、 t_0 でのアブデューサの解仮説中の仮説 η を解としないアブデューサが時区間 $[t_1, t_2]$ 中であっても、 t_0 から t_3 までの一貫性は、 t_0 から t_3 まで分野 η への興味が続く確率 p で与えられる。したがって、 p が十分大きければ t_3 に t_0 の仮説が考慮される。しかし、

*7 “abduct”および後述の“reason”はそれぞれ、“abduction”および“reasoning”、“reasoner”の語尾を除く処理(4.1節参照)を施したキーワードである。

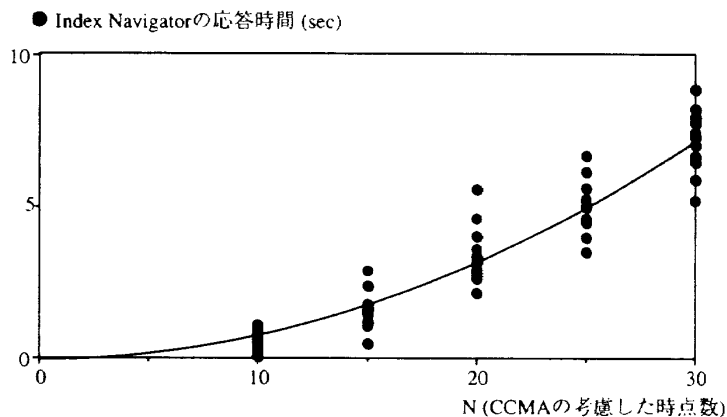


図6 Index Navigatorの応答時間

ベイジアンネットワークの学習データにおいて η の真偽が頻繁に変化していると、 p が小さくなるため t_0 から t_3 への一貫性がほとんど考慮されなくなる。すなわち、学習データでの変化の速さが、今起きている興味の変化と一致しなくても推論結果を左右してしまう。

我々が扱っているのは、学習データでの変化の速さが、実際の変化と一致しないことが多い場合であるから、ユーザの興味の変化があると、ベイジアンネットワークの誤推論はCCMAの誤推論以上に広い時点にわたって多発する。一貫性を強く仮定した他の推論（例えばEC）でも、この性質は同様である。

4・5 Index Navigatorの計算速度

Index Navigatorの中で最も計算時間のかかるプロセスは、CCMAによるユーザの興味理解部である。筆者らは、CCMAにおける各アブデューサの推論エンジンとしてネットワーク化バブル伝播 (Networked Bubble Propagation: NBP と略) 法 [Ohsawa 97a] を用いることにより高速化を図った。

NBP法は、厳密な最適解を求めるCBAがNP完全であるところを、準最適解^{*8}を $O(|H|^2)$ の多項式時間で求めることができる ($|H|$ は仮説数、すなわち文書DB中の分野数)。この様な近似解法を用いた理由は、CCMAにおいてはアブデューサ間の協調によって、解の精度を互いに補い合うことができるからである。

CCMAの計算時間は $O(N^2 \cdot |H|^2)$ のオーダーと予想される [大澤 97b] (N はアブデューサの数、すなわちCCMAが考慮する観測時点数)。 $N = 1, 2, \dots, 30$ に対して、Index Navigatorの応答時間は図6 (計算機は64MB, Pentium-Pro 200MHz, Linux) のようになった。応答時間は平均で $O(N^2)$ となっており、 $|H|$ は N と独立なので $O(N^2|H|^2)$ の予想に一致する。この結

*8 コストがほぼ最小で式(1)～(3)を満たす解仮説。

果から現在のIndex Navigatorでは、システムが5秒以内で応答するように、推論時 (ユーザがキーワードを入力する各時点) の直前の20時点で指定されたキーワードを入力としている (すなわち $N = 20$)。

なお、知識ベースの密度 (分野数とルール数) は、CCMAの協調の効果と、各アブデューサの推論の正確さの双方に影響する。現在の知識ベース (4・1節参照) よりも更に密度が高くなると、より正確な推論結果が得られる可能性もある。しかし、同時に推論時間も増大してしまうので、あまりサイズの大きなデータベースについては今のところIndex Navigatorを用いていない。今後、更に大きな文書DBにIndex Navigatorをサーチエンジンとして用いるには、 $|H|$ の値が大きくなるので更にCCMAの高速化が必要となる。

5. 結 論

Index Navigatorなる文献検索支援ツールを構築した。Index Navigatorがユーザの興味の変化を理解する高い精度により、ユーザは正確に自分の興味をキーワード表現できるようになった。CCMAの性能がベイジアンネットワークを上回ったのは、ユーザの興味が外界の影響を強く受けて変化する為であると考えられる。今後はCCMAの並列化などによる高速化により、Index Navigatorを更に大きな文書DBのサーチエンジンとしても適用可能としたい。

◇ 参 考 文 献 ◇

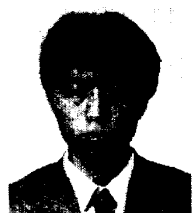
- [Armstrong 95] Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T., WebWatcher: A Learning Apprentice for the World Wide Web, *AAAI Spring Sympo. Series on Information Gathering from Distributed, Heterogeneous Environments* (1995)
- [Balabanovic 95] Balabanovic, M. and Shoham, Y., Learn-

- ing Information Retrieval Agents: Experiments with Automated Web Browsing, *AAAI Spring Sympo. Series on Information Gathering from Distributed, Heterogeneous Environments* (1995)
- [Carver 95] Carver, N. and Lesser, V., The DRESUN Tested for Research in FA/C Distributed Situation Assessment, *Proc. Int'l Conf. Multiagent Systems*, pp.33-40 (1995)
- [Carver 96] Carver, N. and Lesser, V., A Formal Analysis of Solution Quality in FA/C Distributed Sensor Interpretation Systems, *Proc. Int'l Conf. Multiagent Systems*, pp.11-17 (1996)
- [Charniak 94] Charniak, E. and Shimony, S.E., Cost-based abduction and MAP explanation, *Artif. Intell.* Vol.66, pp.345-374 (1994)
- [Fung 95] Fung, R. and Favero, B.D., Applying Bayesian Networks to Information Retrieval, *Com. ACM*, Vol.38, pp.42-48 (1995)
- [Furnas 87] Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T., The Vocabulary Problem in Human-System Communication, *Com. ACM*, Vol.30, pp.964-971 (1987)
- [Kashino 95] Kashino, K., Nakadai, K., Kinoshita, T. and Tanaka, H., Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism, *Proc. IJCAI'95*, pp.158-164 (1995)
- [Krulwich 95] Krulwich, B., Learning User Interests Across Heterogeneous Document Databases, *AAAI Spring Sympo. Series on Information Gathering from Distributed, Heterogeneous Environments* (1995)
- [Ng 90] Ng, H. and Mooney, R., On the Role of Coherence in Abductive Explanation, *Proc. AAAI'90*, pp.337-342 (1990)
- [Ohsawa 97a] Ohsawa, Y. and Ishizuka, M., Networked Bubble Propagation: A Polynomial-time Hypothetical Reasoning Method for Computing Near-optimal Solutions, *Artif. Intell.* Vol.91, pp.131-154 (1997)
- [大澤 97b] 大澤・谷内田, 仮説推論のコストに基づく協調による時系列の理解, *人工知能学会誌* Vol.12, No.4 (1997)
- [Porter 80] Porter, M.F., An Algorithm for Suffix stripping, *Automated Library and Information Systems*, Vol.14, pp.130-137 (1980)
- [Swaminathan 93] Swaminathan, K., Tau: A domain-independent approach to information extraction from natural language documents. DARPA workshop on document management, Palo Alto (1993)

[担当委員: 竹澤寿幸]

著者紹介

大澤 幸生(正会員)は, Vol.13, No.3, p.423 参照.



須川 敦史

1998年大阪大学基礎工学部卒業.

<sugawa@yachi-lab.sys.es.osaka-u.ac.jp>

谷内田 正彦(正会員)は, Vol.13, No.3, p.423 参照.