データベースシステム 110－6
（1996. 10. 24）

情報探索のための自己組織化アプローチ

仲川 亜希　　小西 修

高知大学大学院理学研究科情報科学専攻
高知大学理学部情報科学科

Email:{akinakag,konishi}@is.kochi-u.ac.jp

前回の我々の論文では、与えられた集合の要素間の共出現対に着目したクラスタリングにより、その集合の概念空間を表す方法を、自己組織化マップアルゴリズムを用いて行うことを示した。

World Wide Web(WWW) 上の情報探索では、Yahoo のような WWW ロボットを使った、ユーザの指定するキーワードサーチと Netscape などのブラウザによる情報ブラウジングが行われる。しかし、そこでは、キーワードによる検索集合結果が余りにも大きすぎたり、日々刻々と増進・変化する動的な検索対象世界への対応ができないなど、いくつかの重要な問題をかかえている。

本稿では、我々の自己組織化アプローチを、WWW 上の情報探索に適応し、よりユーザの要求にかなった情報探索法を実現することを示す。WWW ドキュメント集合の中のキーワード対の情報を入力ベクトルとした自己組織化マップ処理により、対象カテゴリがより詳しくクラスタリングされる。そこから自動的に抽出される知識（ルール）に基づいて、ユーザに代わってエージェントが情報探索を行う。

An Self-Organizing Approach to Web Information Search

Aki Nakagawa　　Osamu Konishi

Department of Information science, Kochi University

Email:{akinakag,konishi}@is.kochi-u.ac.jp

In our previous paper, we have described the approach that aims at the relations of term co-occurrence into a given text set, classifies the set of the term pairs using self-organizing map, extracts automatically the conceptual relationship for a cluster, and then displays visually it.

The information search mechanisms provided by the World Wide Web (WWW) are based on keyword searching by WWW robots such as the Yahoo server etc. and information browsing by browsers such as the Netscape etc.. We must however solve some significant problems that a set of keyword search results is too many or we could not enoughly correspond to dynamically increasing and changing information on the Web.

In this paper, our proposed approach attempts to provide an alternative concept-based clustering and searching capbility for WWW servers based on a self-organizing map algorithm and automated knowledge acquisition algorithms. First the content of Web documents is classified by using a clustering algorithm employing the self-organizing feature map. The category hierarchies created could improve Web keyword searching and/or browsing. This process could also be used to develop intelligent agents for more efficient and optimal client based search of relevant Web information.

1 はじめに

WWWのめざましい普及によって、オンライン情報源からの情報の取得が簡単にできるようになった。しかし、それらの使用には巨大な情報空間を探索し、望むものを見つけられないというユーザの問題を伴う場合が多い。WWWドキュメント（ホームページ）などとして提供されるオンライン情報量が増加して、必要とするデータを提供しているサイトの発見や、それらのデータが相互にどの様な関係を持っているのかを把握することが困難になり、インターネット検索を難しくしている。

このような状況において、知識発見技術を使ってデータに埋もれている法則や関係などの有用な情報を検索ユーザにフィードバックして活用することが必要となってきた。[6]

前回の我々の論文では、与えられた集合の要素間の共出現対に着目したクラスタリングにより、その集合の概念空間を表す方法を、自己組織化マップアルゴリズムを用いて行うことを示した。

WWW上の情報探索では、YahooのようなWWWロボットを使った、ユーザの指定するキーワードサーチとNetscapeなどのブラウザによる情報ブラウジングが行われる。しかし、そこでは、キーワード(以降KW)による検索集合結果が余りにも大きすぎたり、日々刻々と増進・変化する動的な検索対象世界への対応ができないなど、いくつかの重要な問題をかかえている。

本研究では、WWWロボットにより収集したWebドキュメント集合に対して、その要素であるKWの共出現語対に着目し、学習したニューラルネットワークから「知識」としてのルールを生成して各クラスタ内の概念関係を抽出する。そうして得られた概念空間を用いて、効果的な探索を支援するシステムについて提案する。

2 WWWドキュメント集合からの共出現語対の抽出

2.1 robotによるWeb情報の収集

ネットワークからのWWWドキュメントの収集は、ロボットによって行われる。ロボットとは、Web上のハイパーテキスト構造を自動的にたどるプログラムである。具体的には、ある一つの文書（ページ）を取得し、次にはそのページの中から参照されているすべてのページを見に行く、ということを再帰的に繰り返していく。

出発点となるURLのリストが与えられると、ロボットはその中から訪問するURLを選び、後は次々に文書を解析しながら新しいURLを取得していく。

ロボットが、文書を認識できればその内容を解析し、データベースに加える。実際にどんなことをするかは、ロボットによって異なるが、HTMLのタイトル部分や、最初の数段落だけを集めるもの、全体を検索して、HTMLの構成に応じた重み付けをしながらすべての単語を集めるものなど様々である。METAや、その他の特別なタグを認識するロボットもある。[3]

このように、パーサ(parser)によって「URL、タイトル、見出し、アンカー文字列」などの要素を解析してKWを解析してKWを抽出し、ドキュメントNoとKW集合からなるデータをデータベースに収納する。

3 共出現語対(co-occurrence term pairs)の抽出

ある文脈において、一つのKWと共に出現するもう一つのKWの間には、概念において何らかの共通の世界(その文脈の主題)を持つと考えられる。[2]そこで、WWWドキュメント集合における共出現語に着目し一つのドキュメントに共出現するKWの対を抽出する。このKW対に頻度情報による順序関係を

持たせることによって概念の階層関係を導入する。

3.1 共出現語関係

KW間の概念の階層関係の情報を得るために同じドキュメント中に共出現するKWとKWの対（共出現語対）とその順序関係を求める。

定義 1 WWW ドキュメント集合 $D = (D_1, D_2, \cdots, D_n)$
ドキュメント D_i=URL、タイトル、見出し、アンカー文字列
WWW ドキュメント集合から抽出されるKWを $\mathrm{TERM_k} = (t_{1k}, t_{2k}, \cdots, t_{nk})$,(t_{ik}はドキュメント D_iの KW) とすると、共出現語対は
$C(\mathrm{TERM_k}, \mathrm{TERM_h})$
$= \{[t_{ik}, t_{ih}] | t_{ik} \in D_i, t_{ih} \in D_i\}$
となる。

定義 2 $C(\mathrm{TERM_k}, \mathrm{TERM_h})$ に重みを付けるために、$\mathrm{TERM_k}$ と $\mathrm{TERM_h}$ の間の距離 (結合度) を次のような関数で与える。
$f(\mathrm{TERM_k}, \mathrm{TERM_h})$

$$= \frac{freq.of C(\mathrm{TERM_k}, \mathrm{TERM_h})}{\sqrt{[freq.of \mathrm{TERM_k} \times freq.of \mathrm{TERM_h}]}}$$

ここで、$freq.of\mathrm{TERM_k} = \sum t_{ik}$
$freq.of\mathrm{TERM_h} = \sum t_{ih}$

定義 3 $C(\mathrm{TERM_k}, \mathrm{TERM_h})$ において、$freq.of\mathrm{TERM_k} > freq.of\mathrm{TERM_h}$ ならば、そのとき $\mathrm{TERM_k}$ は $\mathrm{TERM_h}$ よりも概念の上位関係にあるとする。
$freq.of\mathrm{TERM_k} \geq freq.of\mathrm{TERM_h}$

このように順序を有する共出現語対の二項関係を共出現語関係と呼ぶ。

3.2 共出現語対の抽出アルゴリズム

次にこの定義に基づいた共出現語関係の抽出手順を示す。

共出現語対の抽出では、KWからKWのマトリックス上の組み合わせが考えられドキュメント中の共出現回数を求めるために、$n(n-1) \times m$(ここで、n は一ドキュメント中から抽出されたKW数、m はドキュメント数を表す) 回の計算を必要とする。

step1 2.1で抽出したKWを出現頻度の降順にソートした用語候補リストを準備する。

step2 KW候補リストを得た最初の検索結果のドキュメント集合を対象に、KW候補リストの各KWを検索語とした検索を行なう。

T_1	T_2	f_1	f_2	pair_f	cohesion
learning systm	artificial intelligence	2068	665	197	0.1679
learning systm	backpropagation	2068	154	76	0.1346
⋮			⋮	⋮	⋮

図 1: 共出現語対データの例

入力ベクトル

	artificial intelligence	back -propagation	learning system	neural net	training	...
artificial intelligence	1.000	0.183	0.167	0.350	0.000	...
back -propagation	0.183	1.000	0.134	0.185	0.000	...
learning system	0.167	0.134	1.000	0.655	0.172	...
neural net	0.350	0.185	0.655	1.000	0.142	...
training	0.000	0.000	0.172	0.142	1.000	...
⋮	⋮	⋮	⋮	⋮	⋮	

図 2: 入力パターン例

step3　その検索結果の集合から KW 候補リストと同様に KW を切り出し、頻度の降順にソートする。ここで、検索語となった KW(頻度統計の第 1 位の KW) とそれ以外の KW(第 2 位以降からある頻度以上のものまで) との組み合わせが、共出現語対である。このとき、第 2 位以降の KW の頻度は第 1 位の KW との共出現回数を示している。

step4　KW 候補リストの全ての KW について、**step2,3** を繰り返す。

step5　得られた共出現語対に対して、**定義 2** による結合度を計算する。このようにして得られた共出現語対のデータは図 1に示すような属性をもった関係データベースとして構築される。

4　Kohonen の自己組織化マップによるクラスタリング

WWW ドキュメントをその内容に従って分類するために、Kohonen の自己組織化マップを用いた。

4.1　自己組織化マップ

Kohonen の自己組織化 (特徴) マップ (Self-Organizing (Feature) Map) は、1990 年に T.Kohonen によって提案されたパラダイムであり、ベクトルで表される入力パターン間の位相関係を、学習アルゴリズムにより発見、分類して位相地図を組織化する 2 層のネットワークである。このときベクトルの各成分はパターンの要素に対応している。この結果得られた地図は、ネットワークに与えられたパターン間の自然な関係構造を表している。ネットワークは処理ユニットの入力層と競合層の組み合わせであり、教師なし学習により訓練される。入力パターンは競合層で活性化されるユニットにより分類される。パターン間の類似は競合層のグリッド上の近さの関係に写される。訓練が終了した後、パターン関係やパターングループが競合層で観察される。

Kohonen の自己組織化マップのアルゴリズムは以下の通りである。

自己組織化アルゴリズム

step1 入力パターンを与える。

$E = [e_1, e_2, e_3, \ldots, e_n]$

step2 この入力から競合層の各ユニット i への結合の重みを与える。

$U_i = [u_{i1}, u_{i2}, \ldots, u_{in}]$

step3 その重みが入力パターンと最もよく一致する競合層のユニット c を定める。すなわち、ベクトル E と U_iの間の距離が最小となるものを探す。

$$\| E - U_c \| = \min_j \| E - U_i \| = \sqrt{\sum_j (e_j - u_{ij})^2}$$

step4 このユニット i とその近傍 N_cで重みを調整して一致を増大させる。

$$\Delta u_{ij} = \begin{cases} \alpha(e_j - u_{ij}) & (i \in N_c) \\ 0 & (i \notin N_c) \end{cases}$$

また

$$u_{ij}^{new} = u_{ij}^{old} + \Delta u_{ij}$$

$$\alpha_t = \alpha_0 \left(1 - \frac{t}{T}\right)$$

ここで、αは学習率でその値は訓練が進むにつれて0へと減少していく。また、tは現在の訓練回数であり、Tは行われるべき訓練の全回数である。

step5 学習反復が進むに連れて近傍のサイズと重みの変化の量を次第に減少させる。

4.2 WWW ドキュメント集合からの自己組織化マップの生成

KWの特徴ベクトルを生成する際に、第3章で得られた共出現語関係を用いる。対象となるすべてのホームページから、出現頻度のある程度高い共出現語関係をもつKWを取り出し、これらの共出現語対の結合度をKWの重みとして与える。入力パターン例の一部を図2に示す。こうして得られた特徴ベクトルを用いて Kohonen の自己組織化アルゴリズムを適用して学習を行なう。

4.3 クラスタリング

4.2で述べたように、ニューラルネットワーク技術を用いるとWWWドキュメントをそのドキュメントの内容によって分類された Kohonen マップを生成することができる。こうして得られたマップでは自動的に関連の強いKWが近くにまとめられる。マップ上のKWは、その専門分野の概念体系を表している主要なKWでありこれらの代表的なKWに連なって他の多くのKWがあると考えられる。そこで、これらのKW間の関係からその専門分野の知識構造を把握するために、マップ上のKWをクラスタリングする。

いくつかの出力ノードをひとつのあるクラスタにグループ化し、ルールを定義して概念関係を識別する。クラスタは、それぞれの出力ノードに関するルールの条件文により決定する。すなわち、条件文に同じ属性群を含むルールの出力ノードは同じクラスタにグループ化される。そして、第5章で述べる方法により、これらの概念(クラスタ)は階層化される。

5 クラスタからのルール抽出

ルール抽出のアルゴリズムは、以下の通りである。[7]

ルール抽出のアルゴリズム

step1 すべての入力から競合層のあるユニット b_kへの結合の重みの中で、最大のものを探す。

$W_{max} = (w_{1k}, w_{2k}, \ldots, w_{nk})$

step2 $W_{ik} \geq \beta W_{max}$ となる入力 a_iをすべて選ぶ。ここで β は 0 と 1 の間の定数とする。

step3 **step2** で選んだすべての入力を AND でつなぎ、ルールの条件文とする。例えば、**step2** で選ばれた入力を a_{i1}, a_{i2}, a_{i3} とするとルールは

IF (a_{i1} AND a_{i2} AND a_{i3}) THEN(b_k)

となり

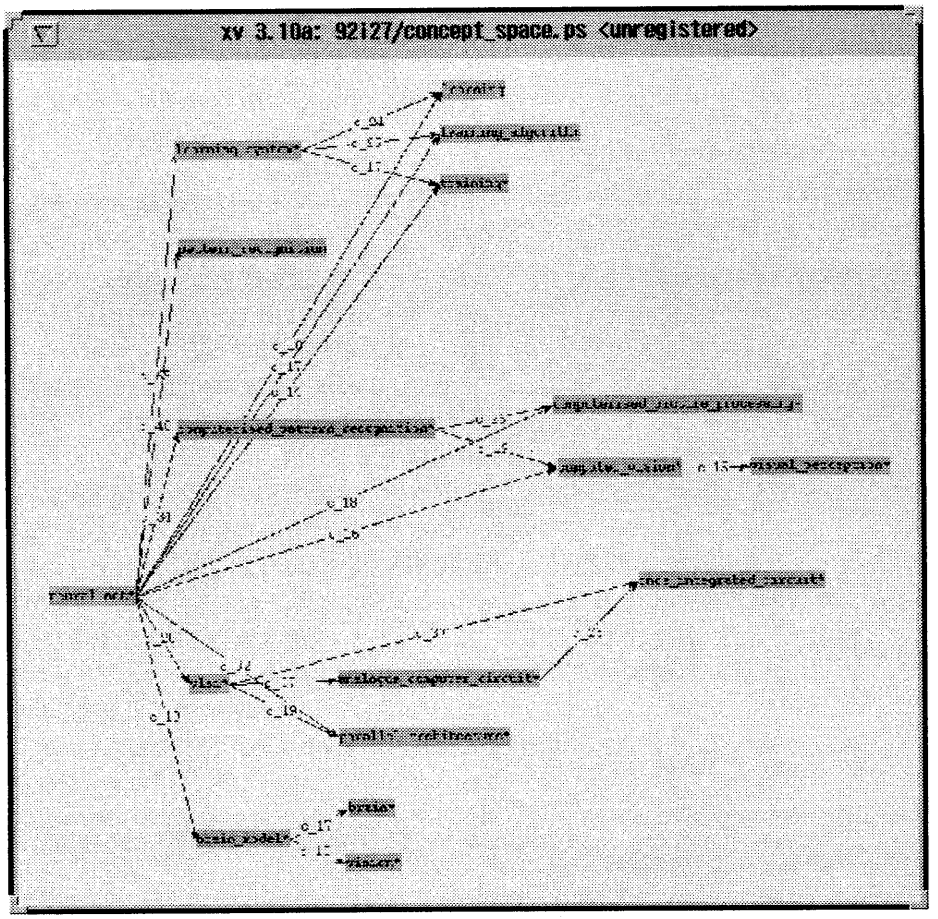

図 3: 概念空間の例

$(a_{i1}$ AND a_{i2} AND $a_{i3}) \Rightarrow (b_k)$
と表す。

step4 **step3** をすべての出力に対して行い、初期ルール集合をつくる。

step5 条件文の中の入力属性が最も少ないルールを選ぶ。

step6 初期ルール集合に、 **step5** で選んだ概念を代入する。

step7 代入がそれ以上できなくなるまで、**step5,6** を繰り返す。

step8 最終ルール集合から、概念階層をつくる。

6 概念空間としてのディスプレイ

特有の主題カテゴリーのホームページ中のKW の共出現語対に着目することによって、それぞれの主題カテゴリーのための概念空間(図 3) を創ることができる。この概念空間は、重要なタームとそのグラフ構造の中で重みのある関係を表現することができ、それは結合

シソーラスに類似している。創造された概念空間は、巨大なインターネットサービスを主題別のカテゴリーとデータベースに分割し、WWWでのKW検索などのユーザの探索を支援することができる。ユーザは、与えられた問題世界の内容全体を概覧することができ、その概念空間中のKWをユーザーが2つ指定すると、それら2つのKWに共通に関連した情報が探索できるような対話型検索が可能となる。また、利用者が検索対象のことを詳しく知らなくてもあるKWに対して関連性の深いKWを選びそれらで絞り込むことによって目的のデータ資源に到達することができるだろう。

7 ユーザ登録キーワードと概念階層に基づく agent(robot)

得られた概念空間はユーザの探索能力を高めるだけでなく、ユーザの嗜好に基づいてインターネット上で最適な情報を求めて広範囲の探索を行う知的エージェントの開発にも用いることができる。

まず最初にユーザがホームページ（あるいはタームペア）を指定し、次に、そのホームページ（それぞれのKW）にリンクされた可能なホームページ（KW）をエージェントがたどって行き、関連のあるすべてのホームページのリストをつくる。そしてエージェントは、ユーザに与えられた最初のホームページ（タームペア）に最も関連の深いホームページを識別する。このように、最初にＵＲＬを指定するだけで、その後の過程はエージェントがユーザの代わりに行い、最も適するホームページに到着することができる。

また、ユーザがコンピューターから離れている時でもユーザの登録したKWに従ってエージェントが自動的に概念階層の更新や、目的のデータ資源などの変化を察知して探索を行うようにする、あるいは定期的に探索を行うようにしておけば、ユーザは、検索にかかる長い時間を待たなくても、いつでも最新の情報を得ることができるようになる。このようにエージェントがユーザの対話型検索の代わりをすることも可能になるであろう。

8 おわりに

本研究では、WWW文献集合の要素であるKWの共出現語対に注目し、自己組織化マップ処理によるクラスタから知識としてルールを抽出し、概念関係を自動抽出して得られた概念空間を用いて、ユーザの探索を効果的に支援するシステムについて提案した。

KWはしばしば、ユーザや背景の違いにより異なった概念として扱われることがあるが、共出現語関係を用いることによってユーザが必要としているデータ資源をより明確に絞り込むことが可能である。また、ユーザが検索対象やKW集合に対して正確な背景知識を持っていなくても、データベースの検索結果の集合や、与えられた問題世界の情報集合から、それらの集合の特徴を表す概念体系を表示し、内容全体を概覧しながら探索を行うことができるので、WWWデータ資源のもつ構造の把握が容易になり、ユーザの知的活動を支援することができる。

参考文献

[1] Hsinchun Chen, Chris Schuffels, Richard Orwig ;Internet Categorization and Search : A Self-Organizing Approach, :Journal Of Visual Communication And Image Representation, vol.7 No.1 March, pp.88-102, 1996.

[2] 小西　修：自動構築型知識に基づく専門用語集形成システム，情報処理学会論文誌,Vol.30,No2,pp179-189,1989.

[3] Martijn Koster; WWW Robot Frequently Asked Questions, 1996. 2. 9, http://info.webcrawler.com/mak/projects/robots/faq.html

津村一昌,「WWW ロボット Q & A」,
http://hml.ec.tmit.ac.jp/robotfaq j.html

[4] 仲川亜希, 小西　修,「自己組織化マップを用いたテキスト情報からの知識獲得」, 情処研報.Vol.96, No.68,pp31-36,96-DBS-109,1996.7.

[5] 仁木和久,田中克巳「ニューラルネットワーク技術の情報検索への適用」:人工知能学会誌,vol.10,NO.1,1995.1.

[6] 西村英樹, 伊藤耕一郎, 河野浩之, 長谷川利治,「重み付き相関ルール導出アルゴリズムによる WWW データ資源の発見」, 第 7 回データ工学ワークショップ (DEWS'96),pp79-84,1996.

[7] S.Sestito & T.S.Dillon
; Automated Knowledge Acquisition, PRENTICE HALL, 1994.

[8] Xia Lin, Dagobert Soergel, Gary Marchionini, ; A Self-organizing Semantic Map for Information Retrieval, : SIGIR, pp262-269. 1991.