特 集

WWWにおけるメタ情報源の獲得

Acquiring Meta Information Resource in the WWW

山田 誠二
YAMADA Seiji
国立情報学研究所
National Institute of Informatics
seiji@nii.ac.jp, http://research.nii.ac.jp/~seiji/

keywords: Meta information resource, interactive document retrieval, information stream, Web community, visualization

1. はじめに

これまでのデータマイニングは，既に与えられている大量のデータから，いかに有用な知識を機械学習を用いて抽出するかを目指し，その知識抽出の方法論をさまざまなアプローチにより研究してきた．しかし，**アクティブマイニング** [元田 05] では，外界からデータが与えられ，それを処理するという受動的なシステムではなく，人間を含むシステムが自ら進んで有用なデータを収集する**アクティブ情報収集**が重要になる．そのため，筆者が研究代表を務める「WWW におけるメタ情報源の獲得」グループ（研究分担者：小野田崇，高間康史，村田剛志）では，WWW で有用な情報の収集を目指し，その収集の対象として，**メタ情報源**という新しいカテゴリである情報を提案し，その収集を実現するために必要な要素技術の検討，それらの要素技術の開発を行ってきた．本稿では，WWW からのメタ情報源の獲得のために，我々の行った研究成果を紹介していく．

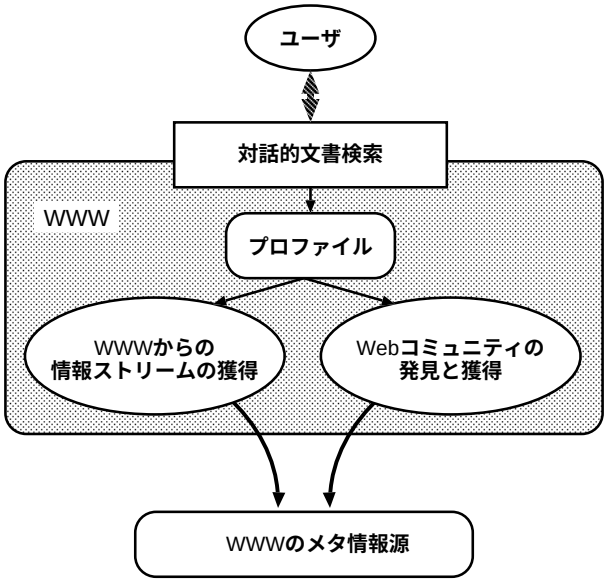

図 1 メタ情報源の獲得

2. WWW のメタ情報源獲得のための技術

アクティブマングの枠組 [元田 05] において，最も能動的に実現されるべき機能は，**情報収集**である．アクティブマイニングは，医療情報データをマイニングの主な対象としている [元田 05] が，我々は，現在最もリッチな情報源である WWW に対しアクティブ情報収集を実現することが，アクティブマイニングの成功に大きく貢献するものと考え，WWW からの情報収集を質的に向上することを目指した．

2·1 WWW のメタ情報源とは何か

これまでの WWW における情報収集 [山田 01] とは，“Web ページの収集” を意味していた．これに対し我々は，構造をもった Web ページの集まりである**メタ情報源**の収集を研究目的とする．メタ情報源とは，単なるばらばらな Web ページの集合でなく，Web ページのコンテンツに加え，Web ページ間の意味のある関係に基づいて取り出される Web ページの時空間的構造を持った集まりを意味する．具体的には，同一のトピックに関連した Web ページからなる集合の時間的な系列である**情報ストリーム**，WWW 上で密に連結している関連 Web ページの集まりである **Web コミュニティ**である．

2·2 アプローチと要素技術

情報収集/検索は，ユーザの興味のある情報を収集/検索することが重要である．メタ情報源の獲得においては，オフラインで行われる情報収集に多くの時間，計算コストがかかる．よって，誤って把握した検索意図による試行錯誤を避けるために，これまでより一層ユーザの検索意図を正確に捉えることが必要になる．このようなユーザの検索意図を把握するためには，ユーザとシステムが対話的にインタラクションを持ちながら，欲しい情報を絞りこんでいく枠組が必須である．このような考えから，我々は，ユーザの検索意図獲得のための枠組として対話的文書検索を採用し，それに分類学習を適用することで

性能向上をはかる．

以上の議論から，WWW のメタ情報源を獲得するために必要な要素技術を，対話的文書検索，WWW からの情報ストリームの獲得，Web コミュニティの発見と獲得，と考え，それらの統合により，効率的かつ効果的なメタ情報源の獲得を実現する．以下に，各技術の概要を示す．

- **対話的文書検索**：従来の適合フィードバックに対し，機械学習の分類学習アルゴリズムを適用して，さらなる検索性能の向上をはかる．分類学習として，記号による判別ルールを学習する関係学習と高次元ベクトルを学習データとするサポートベクターマシンを用い，従来法との比較実験により有効性を示す．
- **WWW からの情報ストリームの獲得**：キーワードを抗体，文書を抗原として免疫ネットワークを構成し，免疫系の特性を有効に活用した情報ストリームの獲得を実現する．静的なトピックを抽出するクラスタリング手法の改良とそのトピックの時間発展を検出，追跡する方法を提案する．
- **Web コミュニティの発見と獲得**：Web ページにおけるリンクの共起を検索エンジンを使って評価することで，密に連結している Web ページの塊である Web コミュニティを発見，獲得する方法を開発する．

メタ情報源の獲得システム全体の構成は，図 1 のようになる．対話的文書検索で適合文書を正確に獲得て，それからユーザの検索意図（プロファイル）を生成し，それを用いて，情報ストリームと Web コミュニティの獲得を的確に行う．

以降では，これらの要素技術に関する我々のグループの研究成果を中心に紹介していく．

3. 対話的文書検索

文書検索において，各文書はそれを特徴づけるベクトルで表現される場合が多い．そのようなベクトルを**文書ベクトル**（特徴ベクトル），文書をベクトルで特徴付ける手法を**ベクトル空間モデル**と呼ぶ．文書ベクトルは，以下の TFIDF 法 [Yates 99] により決定される．今文書集合 D が与えられたとして，その中の各文書を $d_1, \cdots, d_N$ とし，ある文書 d 中における語 t の出現頻度を $tf(t,d)$ で表す．ここで,「語」とは単語から接尾語などを取り除いたものである．また，D において語 t の出現する文書の数を $df(t)$ とする．このとき，文書 d の文書ベクトル v_d は，下式のよう定義される．なお，文書ベクトルの次元は，D の全文書中に含まれる重複しないすべての語で構成される．

$$v_d = (w^d_{t_1}, w^d_{t_2}, \cdots, w^d_{t_n}) \quad w^d_t = tf(t,d) \cdot idf(t)$$

$$idf = \log \frac{N}{df(t)} + 1$$

上式は，ある文書にだけ多く出現し，他の文書にはあまり出現しない語のベクトル値を大きくして，強調している．また，クエリ中の語が対応する次元のベクトル値を 1 に，他の次元のベクトル値を 0 にすることで，クエリ自身も（**クエリベクトル**）と呼ばれる文書ベクトルで記述できる．このベクトル空間モデルにおいて，2 つの文書間の類似度は，それらの文書ベクトルの余弦により定義される．つまり，文書ベクトルの方向が近いものほど，類似していると解釈する．

ベクトル空間モデルに基づく文書検索は，まず文書集合中の文書全てを文書ベクトルで記述する．そして，与えられたクエリをクエリベクトルに変換し，そのクエリベクトルと各文書ベクトルとの類似性を余弦で評価する．最後に，類似度の高い順にソートしたものを適合文書の候補リストとして，ユーザに提示する．ただし，一般にユーザがクエリを正確に記述することは容易ではないため，得られた候補リストは，ユーザが所望していない**非適合文書**をたくさん含んでいる場合が多い．よって，ユーザが欲しい**適合文書**をできるだけ多く見つけることは，一回の検索では難しい．一方，文書が適合文書であるか否かをユーザが判定することは難しくないため，検索結果の文書をユーザに評価してもらえば，検索システムがその評価を利用して，さらに精度の高い検索を行うことが可能になる．このような枠組みが，**適合フィードバック** (relevance feedback) である．適合フィードバックの典型的な手続きを以下に説明する．U と S は，ユーザとシステムが行う処理である．

＜適合フィードバックの手続き＞

(1) U：ユーザがクエリを入力．
(2) S：クエリベクトルを生成し，初期検索を行って結果を得る．
(3) U：ユーザが検索結果の上位 E 個の文書を評価し，適合文書集合 D^+ と非適合文書集合 D^- に分ける．
(4) U：ユーザが十分な適合文書が得られたと判断したら検索終了．不十分な場合は，次へ．
(5) S：D^+ と D^- を使って，クエリベクトルを修正する．
(6) S：修正されたクエリベクトルで，再検索をおこない，Step(3) へ．

一回の検索結果でユーザの評価する文書の数 E は，20〜40 に設定される．Step(5) でのクエリベクトルを修正方法としてよく使われるのが，下の Rocchio の式である [Rocchio 71].

$$Q_{i+1} = Q_i + \frac{1}{|D^+_i|} \sum_{d^+ \in D^+_i} d^+ - \frac{1}{|D^-_i|} \sum_{d^- \in D^-_i} d^-_j$$

上式において，Q_i は i 回目の検索におけるクエリベクトルであり，Q_{i+1} はそれを修正したクエリベクトルであり，d^+, d^- はそれぞれ，個々の適合文書，非適合文書の文書ベクトルである．この式は，直観的には，適合文書に含まれる語の重みはより大きく，非適合文書のそれはより小さくなるようにクエリベクトルを修正している．

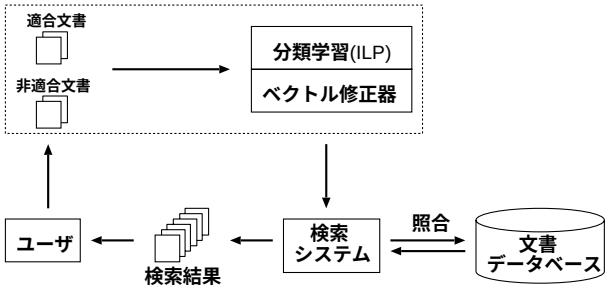

図 2 関係学習による対話的文書検索

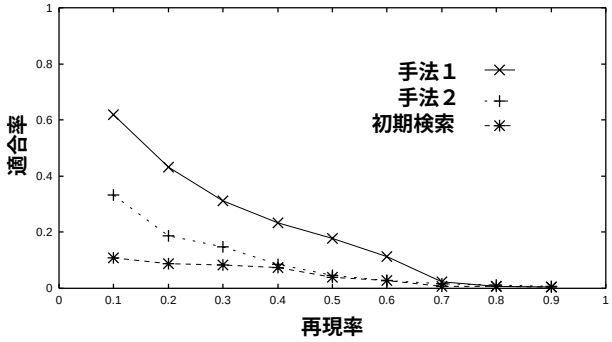

図 3 実験結果

以上が適合フィードバックの基本的枠組みである．以降では，適合フィードバックに分類学習を適用して性能向上を図った 2 つの研究を紹介する．

3·1 関係学習による対話的文書検索

ベクトル空間モデルによる対話的文書検索の利点として，文書をランキングできる，実装が容易といった点が挙げられるが，表現としては限界が存在する．まず，ベクトル空間モデルでは，語の独立性が仮定されているため，語間の近接関係を表現することが難しい．また，同じ適合文書でもそれぞれ注目すべき単語やその組み合わせは異なるが，単一ベクトルでそれらを表現するのは無理がある．これらの情報は，いずれも文書を特徴づけるのに役立つものである．

そこで，ベクトル空間モデルでは記述が難しい情報を，帰納論理プログラミング ILP（Inductive Logic Programming）[古川 01] により文書の判別ルールとして獲得し，そのルールに適合する文書から優先的に順位付けを行うことで，適合文書を効率的に集める研究 [岡部 01] を紹介する．

前述の適合フィードバックのステップ (6) を以下の (6), (7) のステップに置き換えることにより，分類学習アルゴリズムを適合フィードバックに利用することができる．**分類学習** (classification learning) とは，あるクラスに属する正例と属さない負例を与えられ，それらを元に判別ルールや判別関数を帰納的に学習する教師あり機械学習アルゴリズムである．ここでは，ILP をその分類学習アルゴリズムとして用いる．検索ルールの分類学習を組み込んだ対話的文書検索の手続きの概念図を図 2 に示す．

(6) S：適合文書集合 D^+ と非適合文書集合 D^- を用いて，分類学習アルゴリズムにより，判別ルール（判別関数）を作る．

(7) S：全文書を判別ルールを満たす文書集合 A と満たさない文書集合 B に分け，A を上位集合，B を下位集合としたものを検索結果として返し，Step(3) へ．

検索ルールの生成は，適合文書を正例文書，非適合文書を負例文書とした分類学習問題として扱う．検索ルールは，ホーン節で表現する．また，ルールのボディ部は，以下の述語を用いて構成される．near 関係は単語間の近接関係を表すもので，共起関係よりも制約の強い条件となる．

- ap(A, word)： word が文書 A に現れる．
- near(A, word1, word2)： word1 と word2 が文書 A に現れ，順不同で 5 単語以内に近接して存在．

判別ルール学習アルゴリズムの詳細 [岡部 01] は割愛するが，判別ルールを生成するための手続きは，ルールを一つずつ生成し，ルール集合に追加する作業を繰り返す．新しいルールが一つ生成されると，それによって被覆される文書が正例文書集合から取り除かれるので，ルールが生成される度にその集合は小さくなり，最終的に空集合となれば手続きが終了となる．また，ルールは空のボディ部にリテラルを一つずつ，重み付け情報利得に基づき追加していき，負例を一つも含まなくなると完成となる．「米国所得税の脱税者の追跡調査に関する記事」というトピックに対し，実際に学習された判別ルールの例を下に示す．

```
rel(A) :- near(A,charg,hunter),ap(A,count).
rel(A) :- near(A,income,evasion).
rel(A) :- ap(A,evasion),near(A,tax,hunter).
rel(A) :- near(A,evasion,convict),ap(A,illegal).
rel(A) :- near(A,convict,charg).
```

検索対象用の文書データベースとして，文書検索の分野でテストベッドとして広く使われている TREC[Voohees 99] が提供するデータベースの中から英字新聞記事（The Los Angels Times， 約 13 万記事）を使って，評価実験を行っている．その結果の再現率–適合率曲線 [Yates 99] のグラフ（4 回目のフィードバックにおける平均値）を図 3 に示す．手法 1 は提案手法であり，手法 2 は従来の適合フィードバックである．図からわかるように，提案手法がもっとも高い性能を示している

3·2 サポートベクターマシンによる対話的文書検索

本節では，図 2 における分類学習として，高性能な判別関数の学習アルゴリズムとして注目されている**サポートベクターマシン**（以下，SVM）を用いた研究 [Onoda 03] を紹介する．なお，システムの処理手続きは，前節の

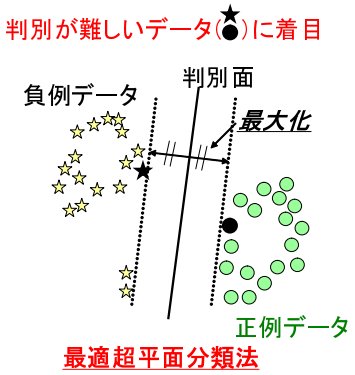

図 4 SVM：サポートベクターマシン

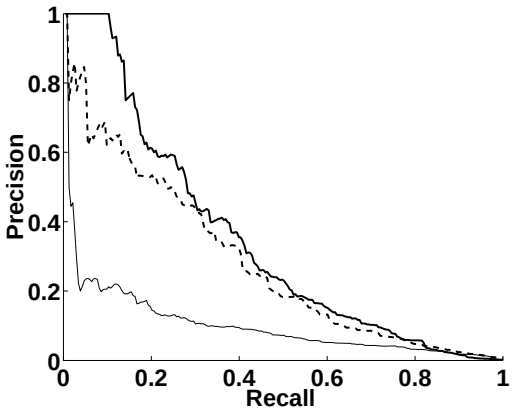

図 5 再現率-適合率曲線:太実線が SVM，点線が Rocchio，細実線がフィードバックなしを表す.

ものと同じである.

SVM[Vapnik 98][小野田 02] は，2 クラスの分類を行う分類学習アルゴリズムであり，与えられた訓練データのうち，サポートベクターと呼ばれるクラスの境界近傍に位置する訓練データと識別面との距離であるマージンを最大化するように判別関数（分離超平面）を構成して分類学習を実現する（図 4）．線形分類ができない場合には，入力空間をより高次の特徴空間に写像して，線形分離を行う（カーネルトリックと呼ばれる）ことで，非線形の問題にも広く適用できる．また，PAC 学習の枠組みから判別関数のクラスの複雑さの測度である VC 次元を用いて，その誤差の上界を見積もることによる，汎化能力の理論的な裏づけがあり，応用面でも様々な分野に利用されている.

SVM を適合フィードバックに適用した結果について述べる．適用データとして，文書検索に関する国際会議 TREC で広く使用されているデータの中の英字新聞記事 (前節で使ったものと同じ) を使用した．このデータには，検索要求文とその要求に適合する文書集合が提供されている．また，ここでの文書ベクトルは非常に語数が多いので，学習サンプルを高次元特徴空間へ写像して分離できるようにする必要がない．そこで，文書ベクトル空間上での線形分離により判別関数を決定した.

SVM の学習に用いる文書数を 20 文書，すなわちユーザが評価を行う文書数を 20 文書とした検索実験を行った．適用結果を図 5 に再現率-適合率曲線で示す．図 5 は，SVM による学習を 4 回行った後の結果を示している．つまり，ユーザが 4 回 20 文書の評価を行った後の結果である．各回にユーザが評価する 20 文書は，SVM が決定した判別関数からの距離が遠く，適合領域に入る上位 20 文書で構成されている．図 5 より，SVM によるフィードバック手法は，フィードバックを行わない場合に加え，従来の Rocchio-based フィードバック手法と比較しても，検索性能が高いことがわかる.

3･3 Web からの情報収集への適用

これまで紹介した対話的文書検索は，素直に WWW の情報収集に適用することが可能である．具体的には，文書データベースを，クエリを入力することにより得られる検索エンジンのヒットリストとすることで，これまで紹介した対話的文書検索の枠組みが，WWW からの情報収集に応用できる．ただし，あくまで文書ベクトルというコンテンツベースで文書を表現するため，ヒットリスト中の Web ページをすべて収集する必要が生じる.

例えば，3･1 節の関係学習による対話的文書検索を，WWW の情報検索，収集に適用した場合，その構成は，図 6 のようになる．詳細は，[岡部 03] を参照されたい.

以上の対話的文書検索に 2 つの分類学習を適用した研究により，従来の適合フィードバックを越えた文書検索性能が達成されている．そして，その対話的文書検索の枠組を WWW からの情報検索に応用することが，可能となる．これにより，メタ情報源の獲得に必要なユーザプロファイル（図 1）を作るために，より高い精度で適合文書，適合 Web ページを得ることができる.

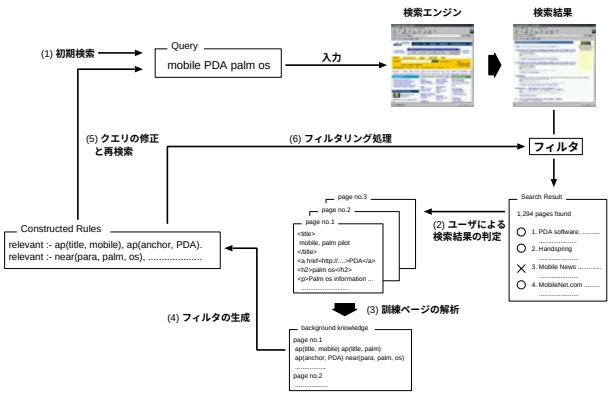

図 6 対話的文書検索による WWW の情報収集，検索

4. WWW からの情報ストリームの獲得

検索結果やオンラインニュース記事集合など，時系列的な関連を持つ文書集合族が Web 上から多く入手可能であることに着目し，これらから話題の流れである**情報ストリーム**を可視化する研究を紹介する．オンラインニュースを始め，Web 上に公開されている情報は，新規の話題や流行に関する情報である**メタ情報源**を多く含んでおり，これらをユーザに提示することで，WWW からユーザがメタ情報源を獲得することを支援することができる．

提案手法は，文書集合毎に可視化を行う際に，過去の可視化結果との対応付けを考慮することで時系列性を考慮する．文書集合毎の可視化は，話題分布をキーワードの空間配置により表現するキーワードマップと文書クラスタリング [Hearst 96] を同時に考慮するために，主要話題に関連しつつ，互いに共起しないキーワード集合を**免疫ネットワーク**モデル [Jerne 73] に基づいて抽出することにより行う [高間 01a]．文書集合間の対応付けは，既使用のランドマークを**免疫記憶細胞**と見なすことにより実現する．

抽出されたランドマークおよび関連キーワードで構成されるキーワードのかたまりは，文書集合中の話題を表現するものであり，これをキーワードマップ上で明確に可視化するために，免疫ネットワーク・メタファを導入する手法についても提案する．キーワードマップ作成の代表的手法であるバネモデル [高杉 99] を基に，ランドマーク関連のバネ長やバネ定数を調整することにより，ランドマークによって表現される話題分布を強調した配置が安定して得ることができる．

4·1 免疫ネットワーク・メタファに基づく情報可視化

キーワードマップ読解の手がかりとなるランドマークとしての性質と，文書クラスタ識別子としての性質を共に満たすキーワードを抽出するために，免疫ネットワークモデルの活性伝播機構が採用されている [高間 01a]．提案する手法では，あるキーワードを共有する文書を**クラスタ**とみなす．このキーワードを**クラスタ識別子**と呼ぶ．

キーワードを**抗体**，文書を**抗原**と見なすことにより，免疫ネットワークモデル (式 (5)–(9)) に基づいてキーワードの活性値を計算し，高活性化したものをランドマークとして抽出する．具体的な処理手順は以下の通りである．ここで，**ネットワークの定常状態**とは，高活性化するキーワード集合が一定となった状態とする．

(1) 文書集合から，出現文書数 DF が TH_2 以上のキーワードを抽出．出現文書集合が等しいキーワードは一つにまとめる．
(2) キーワード間接続強度 (J^b_{ij}) を決定．
(3) キーワード・文書間接続強度 (J^g_{ij}) を決定．
(4) キーワード，文書の活性値計算 X_i, A_i をネットワークが定常状態になるまで繰り返す．

ステップ (2),(3) において，キーワードおよび文書間の接続強度を以下の様に定義する．

キーワード間接続強度（J^b_{ij}）

$$\text{強接続 } (SC) \ \ldots \ CDF_{ij} \geq TH_2 \tag{1}$$
$$\text{弱接続 } (WC) \ \ldots \ 1 \leq CDF_{ij} < TH_2 \tag{2}$$

キーワード・文書間接続強度（J^g_{ij}）

$$\text{強接続 } (SC) \ \ldots \ TF_{ij} \geq TH_1 \tag{3}$$
$$\text{弱接続 } (WC) \ \ldots \ 1 \leq TF_{ij} < TH_1 \tag{4}$$

ここで，CDF_{ij} はキーワード i,j が共起する文書数，TF_{ij} は文書 j 中のキーワード i の出現頻度，SC, WC はそれぞれ，強接続，弱接続の強度を表す．上記条件を満たさない場合オブジェクト間には接続関係はないものとする．

また，ステップ (4) の活性値計算には，本研究では以下に示す数理モデルを使用する [Anderson 93, Neumann 92, Sulzer 93].

$$\frac{dX_i}{dt} = s + X_i(f(h^b_i) - k_b) \tag{5}$$
$$h^b_i = \sum_j J^b_{ij} X_j + \sum_j J^g_{ij} A_j \tag{6}$$
$$\frac{dA_i}{dt} = (r - k_g h^g_i) X_i \tag{7}$$
$$h^g_i = \sum_j J^g_{ji} X_j \tag{8}$$
$$f(h) = p \frac{h}{(h+\theta_1)} \frac{\theta_2}{(h+\theta_2)} \tag{9}$$

ここで，X_i が抗体（キーワード）濃度，A_i は抗原（文書）濃度をそれぞれ表す (初期濃度 $X_i(0), A_i(0)$)．s は抗体の補充率，r は抗原の再生率，k_b, k_g はそれぞれ，抗体，抗原の死滅率である．h^b_i, h^g_i は field と呼ばれ，認識可能な抗原，抗体からの影響は式 (9) より，field の対数を横軸とするベル型の関数により定義される．J^b_{ij} は，抗体 i,j 間の接続強度，J^g_{ij} は抗体 i と抗原 j 間の接続強度を表す．

免疫ネットワークモデルの持つ非線形性により，共起キーワード同士は活性化しあって話題に対応したキーワードの塊を形成すると同時に，活性値が一定以上大きくなると互いに抑制しあうことにより，互いに共起しないキーワードの集合が最終的に高活性化する事が期待できる．

従って，高活性化キーワードを**ランドマーク**とすることにより，キーワードマップ上の話題の分布を理解する手がかりとなると同時に，このキーワードを含む文書単位でクラスタリングを行った場合，クラスタ間のオーバラップを避けることができる．

4·2 免疫記憶細胞モデルの導入

前節で提案したアルゴリズムは，単独の文書集合に適用される．この手法を適用して，時系列的な関連を持つ

文書集合族から話題の流れである**情報ストリーム**を発見，獲得するためには，現在の文書集合を可視化する際に，過去の文書集合から抽出・可視化された話題と類似するものがあれば優先的に抽出・可視化する必要がある．これは，一度ランドマークとして抽出されたキーワードは以降の文書集合において優先的に高活性化する様に優先権を与えることにより実現できる．

実際の免疫システムでは，一度体内に侵入した抗原については免疫記憶細胞が生成され，二度目以降の抗原提示で迅速に反応可能である（二次反応）事に着目し，本稿では，ランドマークとして抽出されたキーワードを以降の処理で免疫記憶細胞と見なす事により，上述の優先権を与える．

免疫細胞モデルについては，(1) 通常細胞よりも低い k_b を与える，あるいは式 (9) で θ_1 を小さく，θ_2 を大きくする，などにより実現可能であり，実験の結果，通常細胞と比較して 6-14 倍，高活性化しやすくなる事が示されている [Takama 01b]．ここでは，(1) を採用して免疫記憶細胞モデルを導入する．

4·3 免疫ネットワーク・メタファに基づくキーワード配置アルゴリズム

文書集合中に含まれる話題分布構造を可視化し，ユーザに提示する手法として，集合中から抽出したキーワードを，文書中における共起関係などに基づき，類似性・関連性の高いほど二次元空間上で近くに配置する**キーワードマップ**が用いられることが多い [高杉 99, 渡部 01]．

免疫ネットワーク・メタファをキーワードマップに導入し，ランドマークに対応する話題を強調した配置を行うことを提案する．ここでは，バネモデル [高杉 99] を，以下の設定を付加することで改良することにより，免疫ネットワーク・メタファを導入する．

- ランドマークに接続しているバネのバネ定数を大きくする．
- ランドマーク間のバネ長を長く設定する．

ランドマーク間の距離を大きくとることにより，ランドマークを中心としたキーワードのかたまりを分離して表示することが可能となる．また，ランドマーク周辺のバネ定数を強くすることにより，ランドマークを中心としたキーワードのかたまりを他よりも優先して形成することが期待できる．さらに，局所最適な配置が得られると言うバネモデルの特徴に関しても，ランドマーク周辺の配置にバイアスを加える形になるため，得られる配置のばらつきが少なくなることも期待できる．

4·4 評 価 実 験

i. 時系列文書集合のクラスタリング結果

前節で提案した情報可視化手法を用いてオンラインニュース記事集合を時系列的に処理した結果について示す．実験には，Yahoo! Japan News (http://yahoo.co.jp/) の「エンターテインメント」カテゴリにおいて，2001 年 9 月 17 日から 21 日の間に公開されたオンラインニュース記事を用いている．

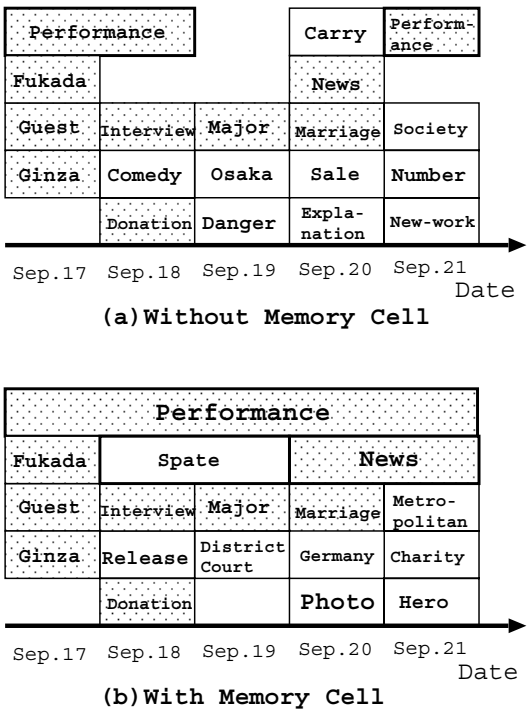

図 7 抽出された話題ストリームの比較 ((a) 記憶細胞なし，(b) 記憶細胞あり)

このオンラインニュース記事を，同じ日に公開された記事集合毎に，提案アルゴリズムで処理を行った結果を図 7 に示す．図 7(a) は免疫記憶細胞モデルを利用せず，各日付毎に独立して処理した場合，(b) は免疫記憶細胞モデルを利用したランドマークに活性化優先権を与えた場合について，生成された各クラスタのランドマークを示している．図 7 中で，免疫記憶細胞の有無によらず，両実験で同様に生成されたクラスタのランドマークについては，網掛けしてある．これより，ランドマークを免疫記憶細胞とすることにより，次回以降のクラスタリングの際に再びランドマークとして抽出されやすくなっていることがわかる．

実験結果の中で,「公演」が全ての記事集合からランドマークとして抽出されている．実験で用いたニュース記事が公開された期間は，米国での同時多発テロ事件直後のため，エンターテインメントカテゴリにおいても関連記事が多数存在しており，その中には，公演の延期やチャリティー公演に関する記事も比較的多かった．提案手法では，多様な話題を発見するために，サイズの大きなクラスタの生成は抑制され，複数のクラスタに分割される傾向にある．そのため，同時多発テロ関係の記事を分割する際に，免疫記憶細胞モデルを導入した場合には一度ランドマークとして抽出された「公演」の観点が再利用される事により，文書集合族を通じた情報ストリームの

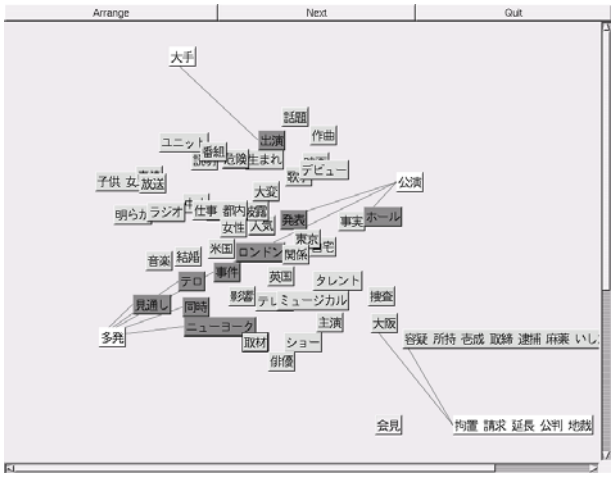

図 8 9/19 のキーワードマップ（免疫記憶細胞あり）

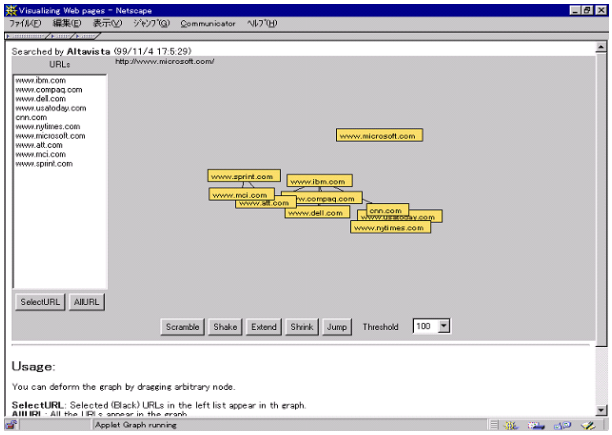

図 9 Jaccard 係数に基づく Web コミュニティの視覚化

一つをとらえることができたと考える．

また，9 月 20 日において両実験により「ニュース」をランドマークとするクラスタが生成されているが，このクラスタに含まれる記事は，同時多発テロ事件を契機に人々がニュースに注目している事を表す，興味深いものであった．

ii. 免疫ネットワーク・メタファに基づくキーワード配置実験

免疫ネットワーク・メタファを導入したキーワードマップの例として，9/19 の文書集合に免疫記憶細胞を用いて提案手法を適用した結果を図 8 に示す．図において，白い矩形のものがランドマーク，濃い色のものが関連キーワードを表す．エッジはランドマークとの強接続のみ示している．これより，関連キーワードの多くはテロに関するものであり，多数の文書に出現する影響で中央に寄った配置となっているものの，各ランドマーク間の距離を大きくとることにより，それぞれが表す話題が明確になっていることがわかる．

5. Web コミュニティの発見と獲得

WWW で最も特徴的な構造とは，Web ページとリンクからなるハイパーテキストであり，そのリンク構造から意味のあるリンク付けされた Web ページのかたまりは，重要なメタ情報源である．そのようなハイパーリンクによって密に結合した関連 Web ページ集合を **Web コミュニティ**と呼び，サーチエンジンからのデータ獲得に基づき Web コミュニィを発見，獲得する研究を紹介する．

主要なサーチエンジンには大量の Web ページが収録されており，巨大で動的な Web を扱う上で重要な資源である．村田は，サーチエンジンを Web データ獲得のための道具とみなし，必要に応じて検索を行なうことで Web コミュニティの視覚化や発見を行なう手法を提案し実験を行なっている [村田 01, 村田 02]．一般に，二つのページ間の関係が密接であるほど，両ページへのハイパーリンクが共起しているような Web ページが数多く存在すると考えられる．提案されている Web ページ間の関連性を視覚化するシステムでは，二つのページの関連性の強さを判定する際，サーチエンジン AltaVista において二つの URL をキーワードとして検索して得られるページ数を，片方の URL をキーワードとして検索して得られるページ数の和で割った値 (Jaccard 係数) を求め，その値が大きいほど URL 間を結ぶ辺の長さが短くなるようなグラフを表示している．この視覚化システムは Web 上に公開しており (http://research.nii.ac.jp/ tmurata/)，実行例を図 9 に示す．サーチエンジンの検索件数を用いたこのような手法は，対象間の関連性を見出す便宜的な手法として有効である．

また，サーチエンジンの検索結果を用いて Web コミュニティを発見する手法として，入力された Web ページ数個を含んでいるような完全 2 部グラフを見出すシステムの構築も行なっている．与えられた Web ページ (centers) 全てにリンクを張っているようなページ集合 (fans) をサーチエンジンのバックリンク検索で求め，その fans から出ているハイパーリンクの中で出現回数が多いものの参照先ページを centers に追加する処理を繰り返すことで Web コミュニティの発見を行なっている．詳細については論文 [村田 01, 村田 02, Murata 03] を参照されたい．

6. ま と め

本稿では，アクティブマイニング実現のために，WWW から収集すべき情報として，従来の Web ページだけではなく，表出されている情報の時空間的な構造をともなった Web ページの集まりであるメタ情報源が重要であることを議論し，具体的に，情報ストリームと Web コミュニティを収集対象に設定した．そして，そのメタ情報源を WWW から獲得するために，ユーザの検索意図獲得の重要性を指摘し，全体のシステム構成に必要な要素技

術として，高性能な対話的文書検索，情報ストリームの抽出，獲得，そして，Web コミュニティの発見を考えた．さらに，それらの技術に関して我々の行ってきた研究である，対話的文書検索，WWW からの情報ストリームの獲得，Web コミュニティの発見と獲得，について紹介した．

これらの技術を統合することで，ユーザとのインタラクションを通じて様々なメタ情報源が WWW から獲得することが可能となる．今後は，そのようなメタ情報源獲得の統合システムの評価が課題である．

謝　　辞

本稿で紹介した研究は，科学研究補助金特定領域研究 (2)「情報洪水時代におけるアクティブマイニングの実現」の支援をうけました．記して感謝いたします．また，本稿執筆に当り，筆者が研究代表者を務める，先の特定領域研究 (2)：研究計画 (2)「WWW におけるメタ情報源の獲得」グループの研究分担者である，小野田崇先生（電力中央研究所），高間康史先生（東京都立科学技術大学），村田剛志先生（国立情報学研究所）に多大なる支援をいただいたことを感謝いたします．

◇ 参　考　文　献 ◇

[Anderson 93] Anderson, R., Neumann, A. U., and Perelson, A. S.: A Cayley Tree Immune Network Model with Antibody Dynamics, *Bulletin of Mathematical Biology*, Vol. 55, No. 6, pp. 1091–1131 (1993)

[古川 01] 古川 康一, 尾崎 知伸, 植野 研：帰納論理プログラミング, 共立出版 (2001)

[Hearst 96] Hearst, M. A. and Pedersen, J. O.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, in *SIGIR'96*, pp. 76–84 (1996)

[Jerne 73] Jerne, N. K.: The Immune System, *Scientific American*, Vol. 229, pp. 52–60 (1973)

[元田 05] 元田 浩：特集「アクティブマイニング」, 人工知能学会論文誌, Vol. ??, No. ?? (2005)

[村田 01] 村田 剛志：参照の共起性に基づく Web コミュニティの発見, 人工知能学会論文誌, Vol. 16, No. 3, pp. 316–323 (2001)

[村田 02] 村田 剛志：ハイパーリンクのグラフ構造に基づく Web コミュニティの 洗練, 人工知能学会論文誌, Vol. 17, No. 3, pp. 322–329 (2002)

[Murata 03] Murata, T.: Visualizing the Structure of Web Communities Based on Data Acquired from a Search Engine, *IEEE Transactions on Industrial Electronics*, Vol. 50, No. 5, pp. 860–866 (2003)

[Neumann 92] Neumann, A. U. and Weisbuch, G.: Dynamics and Topology of Idiotypic Networks, *Bulletin of Mathematical Biology*, Vol. 54, No. 5, pp. 699–726 (1992)

[岡部 01] 岡部 正幸, 山田 誠二：関係学習を用いた対話的文書検索, 人工知能学会誌, Vol. 16, No. 5 (2001)

[Onoda 03] Onoda, T. and Yamada, S.: Relevance feedback with active learning for document retrieval, in *Proceedings of International Joint Conference on Neural Networks*, pp. 1757–1762 (2003)

[Rocchio 71] Rocchio, J. J.: Relevance feedback in information retrieval, in Cliffs, E. ed., *The Smart system – experiments in automatic document processing*, pp. 313–323, Prentice Hall Inc. (1971)

[Sulzer 93] Sulzer, B., Neumann, A. U., Hemmen, van J. L., and Behn, U.: Memory in Idiotypic Networks Due to Competition Between Proliferation and Differentiation, *Bulletin of Mathematical Biology*, Vol. 55, No. 6, pp. 1133–1182 (1993)

[高間 01a] 高間 康史, 廣田 薫：WWW 上の情報収集／可視化のための免疫ネットワークを用いたクラスタリング, 第 46 回人工知能基礎論研究会資料, pp. 61–66 (2001)

[Takama 01b] Takama, Y. and Hirota, K.: Consideration of Memory Cell for Immune Network-based Plastic Clustering method, in *InTech'2001*, pp. 409–414 (2001)

[高杉 99] 高杉 耕一, 國藤 進：スプリングモデルを用いたアイデア触発のための思考支援システムの開発, 人工知能学会誌, Vol. 14, No. 3, pp. 495–503 (1999)

[Vapnik 98] Vapnik, V.: *Statistical Learning Theory*, Wiley (1998)

[Voohees 99] Voohees, E. M. and Harman, D.: Overview of the Seventh Text REtrival Conference(TREC-7), in *Proceedings of the Seventh Text REtrieval Conference*, pp. – (1999)

[渡部 01] 渡部 勇：ビジュアルテキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 226–232 (2001)

[山田 01] 山田 誠二, 村田 剛志, 北村 泰彦：知的 Web 情報システム, 人工知能学会誌, Vol. 16, No. 4, pp. 495–502 (2001)

[Yates 99] Yates, R. B. and Neto, B. R.: *Modern Information Retrieval*, Addison Wesley (1999)

[岡部 03] 岡部 正幸, 山田 誠二：フィルタリングルールの逐次的学習による対話的 Web ページ検索, システム制御情報学会論文誌, Vol. 16, No. 11, pp. 574–582 (2003)

[小野田 02] 小野田 崇：Introduction to Large Margin Classifiers, 人工知能学会誌, Vol. 17, No. 1, pp. 21–30 (2002)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理

著　者　紹　介

山田　誠二（正会員）

1984 年大阪大学基礎工学部卒業．1989 年同大学院博士課程修了．同年大阪大学基礎工学部助手．1991 年同大学産業科学研究所講師．1996 年東京工業大学大学院総合理工学研究科助教授．2002 年国立情報学研究所教授，現在にいたる．工学博士．人工知能，特に，知的 Web，ヒューマンエージェントインタラクションに興味をもつ．情報処理学会，日本ロボット学会，電子情報通信学会，AAAI，IEEE，ACM 各会員．