データベースシステム 119−14
（1999. 7. 21）

部分グラフを基本単位としたWeb文書検索: 単語の出現密度分布の適用

佐野 綾一[†] 松倉 健志[†]
波多野 賢治[†,‡] 田中 克己[†]

本稿では，WWW を Web 文書のリンクにより構成される巨大なグラフとして捉え，検索キーワードを全て含む極小マッチ部分グラフ (Minimal Matching Subgraphs) を検索の基本単位と定義したシステムの評価尺度の提案を行う．この評価尺度は極小マッチ部分グラフのノード，すなわち Web 文書の内部構造を Web 文書内の単語の出現密度分布を用いて抽出しスコアリングを行っている．

Web Document Retrieval based on Minimal Matching Subgraphs as Units and Word Appearance Density

Ryouichi Sano[†] , Takeshi Matsukura[†] , Kenji Hatano[†,‡] and Katsumi Tanaka[†]

A minimal matching subgraph is a minimal subgraph in which all keywords in a query exist, and the end-node contains at least one keyword. In this paper, WWW is regarded as a huge graph of Web document nodes and hyperlink edges, and we propose the evaluation measure of a system that defines a minimal matching subgraph containing all retrieval keywords as the base unit of retrieval. This evaluation measure analyses nodes and inter-node relationship of each Minimal Matching Subgraph (i.e., internal structure of Web document) and performs scoring Minimal Matching Subgraph, using an appearance density distribution of the retrieval keywords.

1. はじめに

WWW (World Wide Web) の目覚ましい発展により，我々はインターネット上から様々な情報を取得できるようになった．しかし，現在の WWW における情報検索では，検索の基本単位として Web 文書単体が用いられているため，ある話題をリンクを用いて複数の Web 文書で記述されている情報は検索が不可能である．

そこで我々は，WWW を Web 文書のリンクにより構成される巨大なグラフと捉え，検索キーワードを全て含む極小マッチ部分グラフ (Minimal Matching Subgraphs) を検索の基本単位と定義したシステムの実装，および評価尺度の提案を行った[2),5)]．しかし，このシステムにおける評価尺度は，極小マッチ部分グラフの各ノード，すなわち Web 文書における検索キーワードの重要度を考慮せずにスコアを決定していたため，検索キーワードと Web 文書に密接な関係が存在したとしても，その関係がスコアに反映されないという問題点を持っていた．

そこで，本稿では Web 文書の内部構造を Web 文書内の検索キーワードの出現密度分布を用いて抽出し，それを反映した極小マッチ部分グラフの評価尺度を新たに提案することで問題点を解決する．

† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
‡ 現在，奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology.

2. 基本的事項および関連研究

2.1 基本的事項

2.1.1 極小マッチ部分グラフ (Minimal Matching Subgraphs)

極小マッチ部分グラフとは，検索キーワードを含まないノード (Web 文書) を端点に持たない，検索キーワード全てを含む極小部分グラフである．

WWW は，Web 文書がリンクでつながった巨大なグラフであると捉えることができるため，

$$WWW = \{U_1, U_2, \cdots, U_l\} \quad (l \geq 1)$$

と定義することができる．ここで，$U_i (i = 1, \cdots, l)$ は Web 文書をノード，文書間をつなぐリンクを有向枝

とする互いに素な連結グラフであり，かつ$U_j \cap U_k = \phi$ $(j,k \in \{1,\cdots,l\},\ j \neq k)$ が成り立っている．

我々の研究における，WWW の検索結果は，図1に示されるように問い合わせ $Q = k_1 \wedge \cdots \wedge k_m (m \geq 1)$ により動的に抽出される．この抽出された検索結果，すなわち極小マッチ部分グラフ $G' = (V', E')$ は以下の条件を満たすものと定義している．

(1) G' は U の部分グラフである．

(2) 各キーワード k_j $(j = 1,\cdots,m)$ に対し，G 内の少なくとも 1 つのノードがこれを含む．

(3) G' のいかなる部分グラフも上記 (1)，(2) を満足しない．

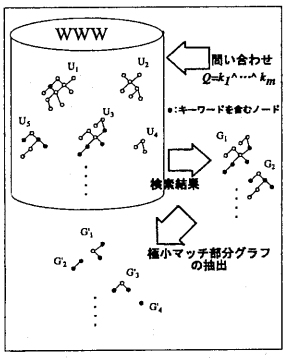

図1 極小マッチ部分グラフの検索モデル

例えば，図2で示されているような，部分グラフ G に対して $Q = k_1 \wedge k_2$ を与えた場合を考える．現在の Web 文書検索システムにおいて Web 文書を基本単位として検索を行った場合，検索結果として返されるのは G'_3, G'_4 だけである．しかし，極小マッチ部分グラフを基本単位として問い合わせを行うことより，リンクによって結びついているノードが極小になるようなグラフ，$G'_1, G'_2, G'_3, G'_4, G'_5$ を得ることができる．

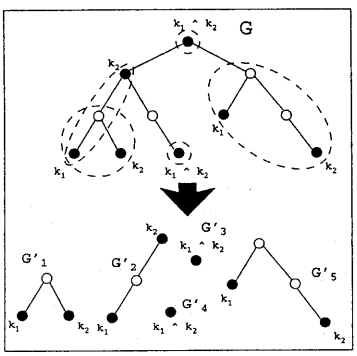

図2 極小マッチ部分グラフの抽出例

2.2 単語の出現密度

単語の出現密度とは，文書中での任意の位置における一定範囲内の単語の頻度と位置的な情報を元にして算出する数値であり，文書中の任意の単語の最重要説明個所の特定に用いられる[7]．

単語の出現密度を計算する際，重み付けにハニング窓関数[8]を使用する．ハニング窓関数 $h_l(i)$ は，窓の幅（重みを与える範囲）を W，窓の中心位置を l とすると，次式であらわされる．

$$h_l(i) = \frac{1}{2}(1 + \cos 2\pi \frac{i-l}{W}) \quad (|i-l| \leq \frac{W}{2})$$

ハニング窓関数を用いた重み付けでは，中心 $(i = 0)$ で 1 となり，中心から離れるにしたがって小さくなる．

以下にハニング窓関数を用いた，単語 t の出現密度の計算方法を示す．

(1) 文書を 1 本の文字列（長さ L 文字）とみなし，文書中での単語のすべての出現位置を調べる．文書の先頭から見て，i 文字目を先頭として単語が出現する場合 $a_t(i) = 1$，そうでない場合 $a_t(i) = 0$ とする．

(2) 位置 0 (文書の先頭) からスタートして，順に各位置をハニング窓の中心位置とし，その中心位置 i に対するキーワードの出現密度 $d_t(i)$ を計算する．ハニング窓の幅を W とすると，中心位置の前後それぞれ $W/2$ の範囲のキーワードの出現を次式によって足し合わせ，出現密度 $d_t(i)$ を求める．

$$d_t(i) = \sum_{j=i-\frac{W}{2}}^{i+\frac{W}{2}} h_i(j) \cdot a_t(j)$$

(ただし，$j < 0$ または $j \geq L$ では $a_t(j) = 0$ とする)

(3) ここで出現密度分布を次式で示すようにその出現密度分布中の最大値で正規化し，最大値を 1 にした $\tilde{d}_t(i)$ を相対出現密度分布と呼ぶ．

$$\tilde{d}_t(i) = \frac{d_t(i)}{\max\limits_{1 \leq j \leq L} d_t(j)}$$

本稿では，この検索キーワードの相対密度分布を利用して，Web 文書の内部構造の解析を行っている．

2.3 関連研究

Web 文書を検索する際に Web 文書を単体として扱うのではなく，1 つ以上の意味的なまとまりをもった Web 文書群を単位として扱う研究は，近年になって様々なものが行われるようになっている．

田島らは，Web 文書を検索する際に検索の単位を Web 文書単体ではなく，Web 文書のリンク構造から静的抽出した「カット」と呼ばれる意味的なまとまり

を検索の単位とすることを提案し，その実装を行っている[4),6)]．これに対し，Li や波多野らは WWW 上のハイパーテキストデータを意味的につながった論理的な文書単位として "information unit" や「極小マッチ部分グラフ」を提案し，それらを基本単位とした Web 文書検索機構の実装を行っている[2),3),5)]．これらの研究は，ユーザが検索機構に問い合わせを行った際に検索結果が動的に得られるという特徴を持っている．しかし，いずれの場合も検索キーワードの重要度の考慮を行っていないため，検索キーワードと Web 文書に密接な関係が存在したとしても，検索キーワードの重要度が検索結果に反映されないという欠点を持っている．

そこで，本稿では Web 文書の内部構造を Web 文書内の検索キーワードの出現密度分布を用いて抽出し，それを反映した評価尺度を提案することで問題点を解決する．

3. Web 文書の検索手法

我々はこれまで，Web 上に存在する情報を一種のグラフとして捉え，ユーザの問い合わせにより，極小マッチ部分グラフを動的に生成する Web 文書検索システムの実装，およびその評価尺度を提案してきた[2),5)]．

しかし，これまでの評価尺度は，極小マッチ部分グラフの各ノード，すなわち Web 文書における検索キーワードの重要度を考慮せずにスコアを決定していたため，検索キーワードと Web 文書に密接な関係が存在したとしても，その関係がスコア上に反映されないという問題点を持っていた．

そこで，本稿では Web 文書の内部構造を Web 文書内の検索キーワードの出現密度分布を用いて抽出し，それを反映した極小マッチ部分グラフの評価尺度を新たに提案することで問題点を解決する．

3.1 文書の内部構造を考慮したスコアリング

極小マッチ部分グラフの検索結果の種類を大別すると図3 のようなものが考えられる．

(1) すべての検索キーワードが 1 つの Web 文書中に出現しているが，その検索キーワード間に全く関連のないもの．

(2) (1) とは異なり，検索キーワード間に関連のあるもの．

(3) 複数 Web 文書に結果が及んでおり，検索キーワードとリンク先の Web 文書との間に関連のないもの．

(4) (3) とは異なり，検索キーワードとリンク先の Web 文書との間に関連の見られるもの．

以上のような違いを考慮して極小マッチ部分グラフのスコアリングを行うためには，検索結果の各文書の内部構造を解析する必要がある．そこで，我々は Web 文書内に出現する検索キーワードの出現密度分布を用いて検索キーワードの locality を算出し，文書の内部構造の解析に利用する方法を提案する．

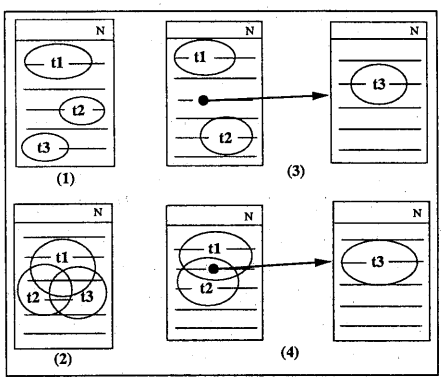

図3　極小マッチ部分グラフの検索結果の種類

単語の locality とは，文書中でその単語の影響が見られるエリアのことである．例えば，図4ように 2 つの単語の locality に重なりが見られる場合，その 2 つの単語は関連している可能性があると考えることができる．すなわち，単語の相対出現密度は単語の locality を表現していると捉えることができる．なぜなら，単語の相対出現密度は任意の位置における特定の語の重要度を表しており，単語の locality とその影響の程度を表現していると考えることが可能だからである．

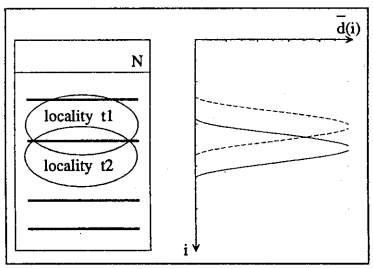

図4　相対出現密度分布と単語の locality

3.1.1 文書中の検索キーワードの関連度の解析

図3の (1), (2) を区別できるスコアリングを行うためには，同一 Web 文書中に出現している 2 つの検索キーワード間の関連度を考慮しなければならない．なぜなら，同一 Web 文書中に 2 つ以上の検索キーワードが出現しているとき，それらはお互いに関連している場合もあれば関連していない場合もあるからである．

文書中での検索キーワード間の関連度は 2.2節で求めた検索キーワードの相対密度分布を利用して算出す

ることができる．相対出現密度分布を利用して，2つの検索キーワード t_1, t_2 の関連度 R を求める式を以下のように定義する (図5参照)．

$$R = r(t_1, t_2) = \max_{1 \le i \le L} \min(\tilde{d}_{t_1}(i), \tilde{d}_{t_2}(i))$$

しかし，文書中で構造的な主題であると考えられる単語 (例えば，文書のタイトルに含まれる単語など) は，文書の一部にしか現れなくとも明らかに文書全体に影響を及ぼし，locality が文書全体にあると考えることができる．よって，検索キーワード t_s が構造的主題である場合については，出現密度の計算結果に関わりなく，以下のようにする．

$$\tilde{d}_{t_s}(i) = 1 \quad (0 \le i \le L)$$

1つの Web 文書中に 3 種類以上の検索キーワードが出てきた場合は，以下の式のように，それらの検索キーワードのすべての組について関連度を算出した後和を取り，その Web 文書内の検索キーワードの関連度 R とする．

$$R = \sum_{t_i, t_j \in T(v)} r(t_i, t_j)$$

(ただし $T(v)$ は文書 v 中に出現する検索キーワードの集合)

こうして文書中の検索キーワード間の関連度 R を算出することによって，同一文書中に同じ種類数だけ検索キーワードが出現した場合の Web 文書間のスコアリングを行うことができる．

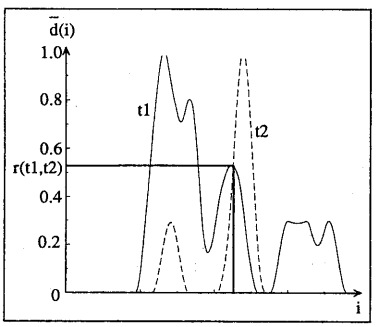

図5　2つの単語間の関連度解析

3.1.2　検索語とリンク先文書との関連度解析

本節では図3の (3), (4) を区別できるスコアリングを行うために，検索キーワードとリンク先文書との間の関連度の解析を行う．

まずリンク元文書において，リンク先文書へのアンカーの出現位置 $i = x$ における検索キーワード t の相対出現密度の値 $\tilde{d}_t(x)$ を求める (図6参照)．この値をリンク元文書中の検索キーワードすべてについて算出し，その最大値を検索キーワードとリンク先文書との関連度 S とする．

$$S = \max_{t_k \in T(v)} \tilde{d}_{t_k}(x)$$

なお，この場合も検索キーワード t_s が文書の構造的主題である場合は x の値に関わりなく，

$$\tilde{d}_{t_s}(x) = 1$$

とする．

以上の方法で，検索キーワードとリンク先文書との間の関連度を解析することにより図3の (3), (4) を区別するスコアリングを行うことができる．

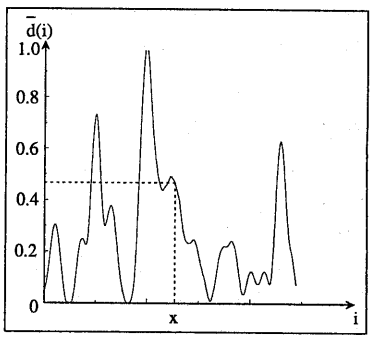

図6　検索キーワードとリンク先文書との関連度の解析

4. システムの実装

出現密度分布を利用した極小マッチ部分グラフのスコアリングは図7のように行われる．

以下に各処理手順を述べる．

4.1　単語の相対出現密度分布の算出部分

検索キーワードの locality を解析するために，検索キーワードの相対出現密度分布の算出を行う．

算出の際の前処理として，HTML::Parse☆ を用いて対象となる Web 文書の構造的主題と考えられる語句を解析し，その中に検索キーワードが出現しているかどうかを調べる．構造的主題と判断された語句中に存在するキーワード t_s に関しては以下の作業を行わず，文書中の相対出現密度の値 $\tilde{d}_{t_s}(i)$ をすべて 1 にする．

(1)　対象となる Web 文書に対して HTML タグの削除処理を行う．これはキーワードの文書中での位置情報が重要な要素を占めるために，実際

☆ libwww-perl[1] の配布している Perl のモジュールファイル．タグなどの解析を行って，HTML 文書をプレーンテキストに変換する．

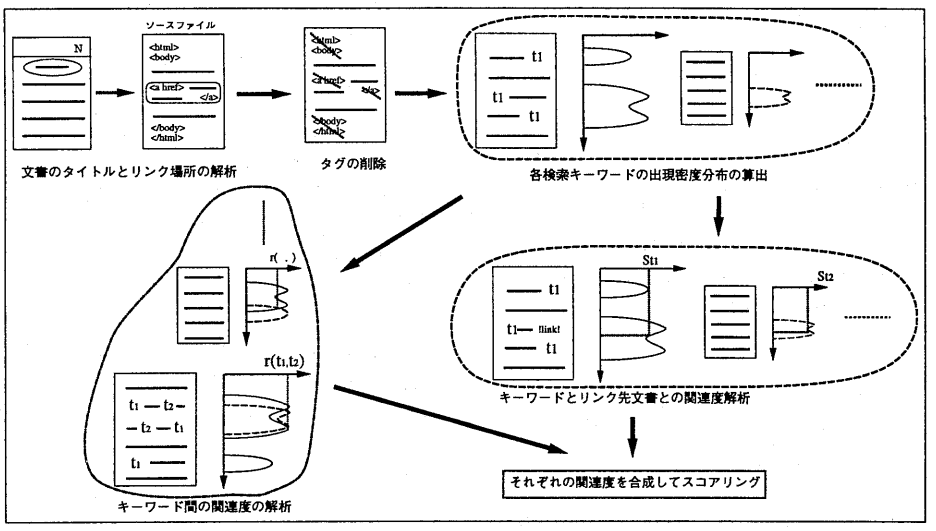

図7　システムの処理の概要図

に表示されない HTML タグは処理の関係上キーワードの正確な位置情報を得るための障害になるためである．

(2)　タグを除去した Web 文書の全文を 1 つの文字列として扱い，文字列の長さ L を測定して，出現密度を求める際に使用するハニング窓関数の窓の幅 W を決定する．また，文書中で使用されているすべての検索キーワードの出現位置を調べる．

(3)　すべての検索キーワード $t_k \in T(v)$ の，文書の先頭 1 文字目から末尾 L 文字目までの出現密度 $d_{t_k}(i)$ を順に求めていく．この際，相対出現密度分布を求めるために，順次それまでに求めた出現密度の値を比較し，そのキーワードの出現密度分布の最大値 $\max d_{t_k}(j)$ を算出しておく．

(4)　算出したそれぞれの検索キーワードの出現密度分布の値をその出現密度分布の最大値で正規化し，相対出現密度分布 $\tilde{d}_{t_k}(i)$ を求める．

4.2　検索キーワード間の関連度解析

次に Web 文書中の検索キーワード間の関連度 R を解析するために，ある 2 つの検索キーワード t_1, t_2 の相対出現密度分布の有値な範囲に重複があるかどうかを調べ，重複があるならば重複部分の最大値を求め，その値をキーワード間の関連度 $r(t_1, t_2)$ とする．

この処理を文書中に含まれる検索キーワード $t_k \in T(v)$ の全組合わせについて行い，求めたそれぞれの値の合計を文書中に含まれる検索キーワードすべての関連度 R とする．

4.3　検索キーワードとリンク先文書間の関連度解析

最後に，検索キーワードとリンク先文書間の関連度 S の解析を行う．解析を行うためには，リンク先文書へのアンカーの位置情報 x が必要である．しかし，HTML タグを削除した後のプレーンテキストではタグ中にあったアンカーのリンク先情報を失っているために，そのリンク先の URL を知ることができない．よって，相対出現密度分布 $\tilde{d}_{t_k}(i)$ の算出処理において Web 文書から HTML タグを削除する以前に，リンク先文書へのアンカーを解析しタグを削除した後もそのアンカーの位置情報 x が判定できるようにする必要がある．そのために Web 文書中のアンカータグの URL 情報が，検索キーワードが含まれている Web 文書への URL と一致するかを調べ，URL が一致すればアンカー文字列の先頭にリンク場所を示すマーカ (例えば {! link !}) を挿入する．その後，検索キーワードの相対出現密度分布 $\tilde{d}_{t_k}(i)$ を求め，タグを削除したプレーンテキスト中でのマーカの出現位置 x を調べる．そして，マーカの出現位置 x における全検索キーワードの相対出現密度の値 $\tilde{d}_{t_k}(x)$ を算出し，その中の最大値を検索キーワードとリンク先文書間の関連度 S とする．

5.　システムの評価

図3の (1), (2) のように単一の Web 文書からなる極小マッチ部分グラフ (すなわち，現在の Web 文書検索システムの AND 検索の結果) と，図3の (3), (4) のように複数の Web 文書からなる極小マッチ部分グラフでは出現密度分布の利用の方法が異なるため，そ

れらを分離して評価を行った.

本評価実験では検索キーワード間の関連度 R と検索キーワードとリンク先文書間の関連度 S のそれぞれの有用性を評価するために次の 2 つの実験を行った.

- 検索エンジンの AND 検索の結果に対して関連度 R を算出しスコアが 0 の Web 文書を除去した場合の適合率の評価実験.
- 極小マッチ部分グラフのノードが 2 つの場合の関連度 S を算出しスコアが 0 の極小マッチ部分グラフを除去した場合の適合率の評価実験.

以下に評価実験の結果を表1に示す.

表1　適合率の変化

	除去前	除去後
AND 検索の結果	33.9	40.1
極小マッチ部分グラフ	30.7	36.0

表1を見ると適合率はどちらの場合も上昇してはいるが, 出現密度分布を利用する方法の有用性を証明できているとは言い難い. しかし, 今回の実験では極小マッチ部分グラフの関連度のスコアが完全に 0 のものだけを除去しているため, 関連度のスコアにある閾値を設定し極小マッチ部分グラフの妥当性を評価する必要があると思われる. さらに, 適合率の大きな上昇が見られない理由として, どちらの場合も単語の出現頻度などを用いた適合文書中の検索キーワード自体の重要度を考慮していないため, 今後, 文書中での検索キーワード自体の重要度を測定し, 検索キーワード間の関連度, 検索キーワードとリンク先文書間の関連度と併せてスコアリングを行うことで, これら問題点を解決していく必要があると思われる.

6. おわりに

本稿では, WWW を Web 文書のリンクにより構成される巨大なグラフとして捉え, 検索キーワードを全て含む極小マッチ部分グラフを検索の基本単位と定義したシステムの評価尺度の提案を行った. この評価尺度では極小マッチ部分グラフのノード, すなわち Web 文書の内部構造を Web 文書内の単語の出現密度分布を用いて抽出することで, 検索キーワードの出現位置を考慮したをスコアリングが行うことが可能となった.

今後の課題としては以下の事項が考えられる.

- 適合率を上昇させるためのスコアの閾値の設定.
- 文書の特徴づけのための検索キーワードの出現頻度やタグ情報の付加.
- 極小マッチ部分グラフが単一の Web文書からなる場合と複数の Web 文書からなる場合における出現密度分布を利用したスコアリングの方法が異なるため, この 2 つの場合の関連度 R と S を統合したランキングの方法. 例えば W_1R+W_2S (W_1, W_2 は重み) のようにそれぞれの関連度の和を取る方法など.

謝　辞

本稿をまとめるにあたり, 有益な御助言と御教示を賜りました神戸大学工学部情報知能工学科の田島敬史助手に謹んで感謝の意を表します.

また, 本研究の一部は, 日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」(プロジェクト番号 JSPS-RFTF97P00501) および, 文部省科学研究費重点領域研究「高度データベース (No.275)」(課題番号 08244103) による. ここに記して謝意を表します.

参考文献

1) http://www.linpro.no/lwp/.
2) Kenji Hatano, Ryouichi Sano, Yiwei Duan, and Katsumi Tanaka. An Interactive Classification of Web Documents by Self-organizing Maps and Search Engines. In *Proc. of the 6th International Confarence on Database Systems for Advanced Applications (DASFAA'99)*, pp. 35–42. World Scientific, Apr. 1999.
3) W. Li and Y. Wu. Query Relaxation by Structure for Web Document Retrieval with Progressive Processing (Extend Abstract. In *Proc. of Advanced Database Symposium'98 (ADBS'98)*, pp. 19–25, Dec. 1998.
4) Keishi Tajima, Yoshiaki Mizuuchi, Masatsugu Kitagawa, and Katsumi Tanaka. Cut as a querying unit for WWW, Netnews, and E-mail. In *Proc. of ACM Hypertext*, pages 235–244, Jun. 1998.
5) 波多野賢治, 佐野綾一, 段一為, 田中克己. 自己組織化マップと検索エンジンを用いた Web 文書の分類ビュー機構. 情報処理学会論文誌: データベース, Vol. 40, No. 1, pp. 933–942, 1999年1月.
6) 水内祥晃, 田島敬史, 田中克己. グラフ構造を持つデータのカット分割に基づく検索. 情報処理学会研究報告, 97-DBS-113-47, pp. 281–286, 1997年7月.
7) 黒橋禎夫, 白木伸征, 長尾眞. 単語の出現密度分布を用いた語の重要説明箇所の特定. 情報処理学会論文誌, Vol. 38, No. 4, pp. 845–854, 1997年1月.
8) 長尾眞. パターン情報処理. コロナ社, 1983.