特　集　「Webシステムにおける情報獲得支援技術」

Webサーチエンジン技術の高度化

Advanced Web Search Engine Technologies

風間　一洋　NTT未来ねっと研究所
Kazuhiro Kazama　NTT Network Innovation Laboratories.
kazama@ingrid.core.ntt.co.jp

原田　昌紀　（同　上）
Masanori Harada　harada@ingrid.core.ntt.co.jp

Keywords: search engine, distributed information retrieval, P2P, link connectivity analysis, HITS, PageRank, anchor text, behavior analysis.

1. はじめに

Web情報検索システムには，一般の情報検索システムにはないようないくつかの課題が存在する．

(1) インターネット上のWebページの総数は指数関数的に増大しており，2000年1月の段階で約17.8億URLであると推測されている[山名00]．そこで，大規模サーチエンジンでは，処理可能なデータ容量の拡大と，処理速度の向上が必要になる．

(2) Webページは新聞記事などとは違い，いったん公開された後も動的・不定期に更新される．このために，検索用の索引との同期が取りにくく，検索結果中に現れる情報がすでに存在しない，または別のサーバに移動していることも多い．

(3) Webページは，ハイパーテキストとしてさまざまな文書構造をもっているために，明確な文書集合を切り出すことが難しい．ページ単位で処理すると，同一文書内の異なるページが多量に表示されてしまうことも多い．

(4) Web上には，あらゆる分野の情報が，さまざまな言語やファイル形式で公開されている．また，新聞記事などとは対照的に，用語や文体が統一されていないうえに，綴り間違いも多く，選択する検索語によっては検索結果を絞り込むときに除外され，目的の情報を見つけ出せないこともある．

(5) 従来の情報検索システムでは，検索技能をもつ専門家を利用者として想定していた．しかし，サーチエンジンの大部分の利用者は，利用方法を読むことも少なく，有用な論理演算子やオプション機能を使わないだけでなく，ごくわずかの検索語しか使用しない傾向がある．例えば，SilversteinらのAlta Vistaのログ解析結果では1～3語（平均2.35語），ODINでは日本語の複合語が空白で分割されない特徴を持つためにさらに少なく，1～2語（平均1.42語）である[Silverstein 98, ODIN]．このために，利用者の真の要求とは異なる順序で大量の検索結果が出力されてしまうことが多い．

(6) 自分のWebページが検索結果の上位に表示される機会を増やすために，例えばWebページの背景色と同じ色でさまざまな検索語を記述するようなスパム行為がさかんに行われている．多くの利用者が閲覧を望まないような内容のWebページに多いこともあり，何らかの対策が必要である．

サーチエンジンにおいては，既存のIR（Information Retrieval）分野で培われてきた技術をベースにするだけでなく，このような課題に対処するために独自の発展をとげつつある．

そこで，本稿では，まずWeb情報検索システムをアーキテクチャの面から整理し，その発展形である並行化，分散化について述べる．続いて，ネットワーク上の情報流通にサーチエンジンがどのように寄与しているかを分析し，具体的にハイパーリンク情報と検索履歴を用いたWeb情報検索に特化した技術や，将来的にWeb情報検索の高度化に寄与すると思われる技術を解説する．

2. Web情報検索システムのアーキテクチャ

2·1 ロボット型とディレクトリ型

Webサーチエンジンは Web情報空間の情報を検索するシステムであり，ロボット型サーチエンジンとディレクトリ型サーチエンジンに大きく分類できる．

ロボット型サーチエンジンでは，Webロボットと呼ばれるプログラムを用いてWebサーバから収集した情報に対する全文検索を提供する．情報の網羅性が比較的高い反面，膨大な検索結果のなかから必要とする情報を見つけ出すのが難しい．さらに，ロボットによる収集に時間がかかるので，検索結果と実際の情報のくい違いが

生じやすい．

ディレクトリ型サーチエンジンでは，Webページの情報をカテゴリ別に分類して提供し，利用者はクリックするだけで使用できる．一般的にこの分類作業は人手で行われるので，ロボット型サーチエンジンの検索結果より内容が整理されているが，登録されるWebページの情報は，エディタが選んだWebページか，登録申請されたWebページに限られるので，網羅性が低くなりやすい．同時に提供されるキーワード検索も，登録されたWebページの要約しか検索対象にならないので，目的の情報が登録されていても検索できるとは限らない．

2·2 並列化と分散化

Webの膨大な情報量に対処するために，並列化に基づく処理の高速化と，分散化に基づく情報管理コストの削減の二つのアプローチが行われている．

§1 並列化

並列化は，情報収集や索引作成，検索などの処理を細分化し，マルチプロセッサマシンや計算機クラスタなどのハードウェア上で同時に実行するアプローチである．サーチエンジンの場合には，Webロボットによる情報収集，索引作成，検索などのさまざまな段階で行われており，処理能力が不足したときには，CPUモジュールや計算機を追加することで対処している．

ただし，サーチエンジンの詳細なハードウェア構成は，必ずしも公開されているわけではない．そこで，まず過去の例をあげれば，AltaVistaは1997年の時点で検索に用いるNI2データベース用に10 CPUのAlpha Server 8400を11台組み合わせて使用していた [野田1997]．また，HotBotではUCB（University of Berkeley）のNOW（Network of Workstation）プロジェクトが開発したワークステーションを高速なネットワークで接続した計算機クラスタを使用していたが，1998年の時点では，50～70台のワークステーションで構成されていた [NOW]．

現在は，扱う情報量が膨大なロボット型サーチエンジンにおいては，マルチプロセッサマシンより，スケーラビリティと性能，費用などの点で有利な計算機クラスタのほうが広く用いられている [赤峯 00, InfoNavigator]．例えば，Googleでは，2000年5月の時点でLinuxサーバを4 000台以上用いていると発表されている [RedHat 00]．

§2 分散化

分散化は，複数のサーバが局所的に管理している情報を，統一的に処理できるようにすることである．情報を分散して保持すれば，サーバの管理コストを減らせるだけでなく，特定のサーバに障害が発生しても，システム全体のダウンを回避することができる．特に，Webロボットを用いてWeb情報空間から大量の情報を収集する代わりに，サーバで局所的に情報を管理すれば，ロボット型サーチエンジンで発生する現実の情報と索引の間の不一致を解消したり，検索される情報を動的に変更することなども可能になる．

分散化は，階層化された制御構造をもつ場合と，フラットな制御構造をもつ場合の二つに大きく分類できる．

前者の例として，複数のサーバに検索要求をブロードキャストし，その結果を統合するメタサーチエンジンがある．これは，MacOSのSherlockのようにクライアント側で検索結果を統合するタイプと，MetaCrawlerなどはサーバ側で検索結果を統合するタイプの二つに大きく分類できる [Selberg 95]．メタサーチエンジンは，既存のサーチエンジンやサイト内検索をそのまま利用でき，結果を統合する処理が比較的軽いなどの利点があるために広く行われている．しかし，選択できるサーバ数や情報量が多くなると，うまく検索結果を絞り込むことができなかったり，目的の情報をもたないサーバにまで問い合わせて余計な時間がかかることになる．そこで，さまざまな専門分野をもつ多くの情報検索サービスのなかから，目的に応じたサービスだけを自動的に選択する方法の研究が行われている [Gravano 99, Sugiura 00]．

ほかに，各検索サーバがもっている情報の一部を上位のサーバが保持することで，実際に検索要求を出すサーバを絞り込む方法の研究も行われている．例えば，Harvestでは，Gathererが収集した文書のキーワードやメタ情報をSOIF（Summary Object Interchange Format）を用いてBrokerに転送するが，Brokerはさらにほかの Brokerに転送することができる [Bowman 94]．ほかに，CSE（Cooperative Search Engine）では，局所メタサーチエンジンが位置サーバに「どのサーチエンジンが情報をもっているか」という forwarding knowledgeを転送し，検索語をもっているサーバに絞り込むことができる [上原 00]．

これに対して，後者は，インターネット上の複数のサーバが協調して情報を検索するアプローチである．この方法では，集中管理する部分が存在しないので耐障害性が高いだけでなく，自律的に動作するので管理コストが低く，大規模分散システムに特に適している．例えば，Gnutellaのような P2P（Pear To Pear）型のファイル検索システムでは，プログラムはまず指定された数のマシンに対して検索要求を出し，各マシンはTTL（Time To Live）が指定された最大ホップ数を越えない範囲でそれをルーティングし，該当するファイルが見つかった場合には，経路を逆順にたどって結果が伝えられる [Gnutella]．同様のサーチエンジンとしてはInfraSearchが存在したが，Sun Microsystems社に買収された後に Projext JXTAに吸収されている [InfraSearch, JXTA]．この方法は，小規模で動的なコミュニティを素早く形成するのに適しているが，通信量が膨大になりやすいので，広範な検索には適していない．

われわれの Ingrid では，まず Web リソースからそれ自身のメタ情報を記述する RP（Resource Profile）を抽出する．さらに，インターネット上に配置した複数の FI サーバ（Forward Information Server）上に，RP をノードとして，できるかぎり多くのキーワードを共有するノード間にリンクを設定したグラフ構造を作成する．特に，このグラフ構造のある部分グラフの任意の 2 ノード間に内部のみを通るパスが存在する場合に，その部分グラフをクラスタと呼ぶ．Ingrid では，RP に含まれるすべてのキーワードの組合せに対して，それをもつ RP の集合がクラスタを成すという条件を満足しながら，できるかぎり粗になるようなグラフ構造を作成し，これを Ingrid トポロジと呼ぶ [Francis 95]．クライアントは，Ingrid トポロジ上を並列に探索して目的のリソースを発見する．実際には，さらにサーバ側とクライアント側で FI をキャッシュして再利用することで，より効率的な探索を行っている．

後継の Pinot では，Ingrid の最初の FI サーバを見つけるための専用サーバの廃止，Ingrid トポロジの耐障害性の向上，ODIN や ODIN Directory に用いられている全文検索エンジン Jerky との統合などが行われている [原田 00, 佐藤 99]．さらに，静的に抽出された RP を用いずに，利用者の検索要求に基づいて動的に Ingrid トポロジを作成し，利用者の使用に応じて自律的に変化させる実験も行っている．

ただし，Gnutella と Ingrid のどちらの方法も，プロセス間通信量の爆発を回避するために，ネットワーク上のどれかのサーバ上にある情報が存在する場合でも，それが必ず検索できることは保証しない．分散システムの規模が拡大した場合に，単純な近接性に基づく Gnutella ではこの点が早期に破綻するが，このような場合に Ingrid トポロジは比較的良好なスケール特性を持つように定義されている．

3. Web における社会性の利用

Web 情報空間に関する基礎的な研究としては，インターネット上に公開されている Web ページ総数の推定や [Bharat 98a, Lawrence 98, Lawrence 99, 山名 00]，Web のグラフ構造の分析 [Albert 99, Broder 00] などが行われている．

これらの分析結果が示すように，Web の普及と共に人々はネットワーク上で情報公開や情報交換，情報入手，物品の購入などさまざまな行動が可能になり，ネットワーク上の社会的な活動の比率が高くなってきている．

これに伴って，図 1 に示すように，ある人が Web 情報空間に情報を公開し，別の人がその情報を閲覧するという情報流通が発生するが，サーチエンジンは，この一部として組み込まれていると考えられる．そこで，Web 情報空間の構造情報と，サーチエンジンの利用者の行動情報を分析・利用して，高度なサーチエンジンを実現するための研究が盛んに行われている．

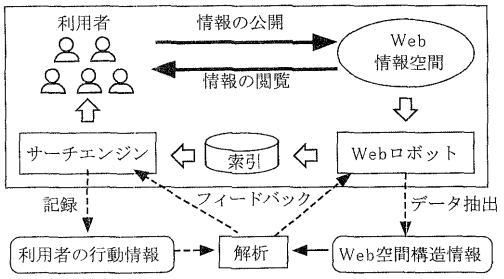

図 1　情報流通とサーチエンジン

本稿では，特にハイパーリンク情報をサーチエンジンのランキング改善やグループ化に適用する方法と，検索履歴を解析して検索支援に適用する方法について解説する．

4. ハイパーリンク情報

一般に，サーチエンジンの利用者が使用する少数のキーワードだけから，どのような情報を探そうとしているかを推測するのは困難である．そこで，Web ページの一般的な価値や関連性を推測して，その推測に基づいて情報を提示する方法が考えられる．

4·1　リンク接続解析

Web で公開されている文書の大部分は HTML で記述されたハイパーテキストであるが，このリンク状況は Web ページの内容や，ほかの人からの評価に応じて大きく変化する．

例えば，文書 A が文書 B にリンクした場合には，文書 A の著者は文書 A と文書 B の間に何らかの関連があると考えていると推測できる．さらに，文書 B をリンクしたということは，文書 B の情報に価値を認めたということを意味するので，文書 B が多くの人からリンクされるほど重要であると推測できる．つまり，リンクすることは，多数の Web ページ作成者による一種の協調フィルタリングであると考えることができる [Goldberg 92]．

この仮定に基づいて，ハイパーリンクの構造を解析して，Web 文書の Web 情報空間内における関連性や重要性を求める手法が数多く提案されている．これらを，リンク接続解析（Link Connectivity Analysis）と呼び，検索結果のランキング精度の改善や，関連情報の発見に利用されている [山田 01]．この手法は，次の 4 種類に大きく分類できる．

§ 1　Link Popularity

Link Popularity は，数多くリンクされている Web ページほど重要だとみなす手法である．例えば，検索された Web ページの被リンク数や，検索された Web ページが存在するサーバ全体の総被リンク数の多い検索結果

のスコアを高くする手法が使われている [SEW].

§2 HITS

Kleinberg らが提案した **HITS**（Hypertext Induced Topic Search）は，ある特定のトピックに関する情報源であるオーソリティ（authority）と，オーソリティへのハイパーリンクの集合であるハブ（hub）という2種類の Web ページに対して，良いオーソリティは多くの良いハブからリンクされ，良いハブは多くの重要なオーソリティをリンクするという相互依存する関係を求めることで，検索結果の質を改善する手法である [Kleinberg 99].

§3 PageRank

Page らが提案した **PageRank** は，多くの良質な Web ページからリンクされている Web ページは，良質な情報源であると考え，Web のリンクをランダムに辿る "random surfer" 行動モデルに基づいて，Web ページが閲覧される確率を計算して得られる Web ページの重要度である [Brin 98, Google].

§4 Cocitation アルゴリズム

Dean らが提案した **Cocitation** アルゴリズムは，何らかの関連をもつ Web ページをリンク解析で発見する手法として，二つの Web ページを同時に引用している Web ページ群を求めて，その数により関連性が高い Web ページを発見する手法である [Dean 99].

4·2 トピックドリフト問題

これらの手法を用いて Web ページ間の重要性や関連性を解析して利用する場合に，検索語と関連のないと思われる Web ページが検索結果の上位にランキングされたり，関連するページとして選択されることがある．これをトピックドリフト問題（topic drift problem）と呼び，次のような原因で発生する [Bharat 98b].

- リンクの意図を考慮しないで，一括してハイパーリンク情報を処理すると，一般的なトピックや著名サイトほど高く評価されやすくなる．
- Web 文書のリンク構造は多種多様であり，例えばプログラムが機械的に生成した大量のハイパーリンクの影響を受けることがある．
- 検索質問とは独立にハイパーリンク情報を処理すると，複数の検索語を指定した場合に，一般的なトピックを示す検索語に偏りやすい．

例えば，HITS や PageRank を単独で用いた場合は，リンク元の信頼度が反映されるのでスパムに強い反面，どのような意図でリンクされているかはいっさい考慮しないので，一般的なトピックや著名なサーバの Web ページに偏りがちであり，常に適切な結果が得られるとはかぎらない [Dean 99, 高野 00]．そこで，この問題を改善するために，次のようなアプローチが試みられている．

§1 ハイパーリンクの隣接関係の利用

同一 Web ページ内にはさまざまな意図に基づいたハイパーリンクが混在しているが，近隣に配置されるハイパーリンクは同じトピックを示している可能性が高い．そこで，Dean らによる Companion と Cocitation では，注目するハイパーリンクの近隣のハイパーリンクだけに HITS を適用している [Dean 99]．さらに，豊田による Companion+では，近いリンクほど高く評価することで，より良い結果が得られるように改良している [豊田 00].

§2 アンカーテキストの利用

ハイパーリンクはアンカー（anchor）と呼ぶ二つの端をもつが，特にソースアンカー中のテキストをアンカーテキスト（anchor text）と呼ぶ．アンカーテキストはリンク先の内容や意図を比較的的確に反映することが知られているので，これを利用すれば検索目的に合致したハイパーリンクを選択することができる．

リンク元のアンカーテキストをリンク先のテキストの一部とみなして検索する方法は，検索語と一致するハイパーリンクだけを選別してから，その Link Popularity を計算してスコアに反映することとほぼ同等であり，より検索目的に合致した検索結果を得ることができる．

ただし，リンク元の信頼度は考慮されないのでスパム攻撃などに弱いが，Google では PageRank を組み合わせることで欠点を補っている [Brin 98].

ODIN では，本文，サーバ内部のリンクとサーバ外部へのリンクを区別し，異なる重みを設定することで，その Web 文書の作者以外の人のリンクを優先的に評価し，より協調フィルタリングに近い特性を実現している [風間 00a]．また，Link Popularity を用いて信頼できるオーソリティからより多く情報を収集するように Web ロボットを制御しているので，リンク元の信頼度の問題は，それらのオーソリティからのハイパーリンクが優先的に扱われることで改善される．

ほかに，Chakrabarti らは，ある Web ページを参照しているソースアンカー周辺のテキスト中のマッチした文字列数をスコアに加算することで，HITS を改善する手法を提案している [Chakrabarti 98].

§3 被参照範囲の制限

Companion や，Companion+，Cocitation では，複数の Web ページに関連した Web ページを発見させる場合に，一部の被参照数が大きい Web ページの影響が支配的になることが多い．

そこで，われわれは多くの異なるシードと参照共起関係にあるほど関連度を高くする MultiCocitation を使用している [原田 01].

§4 ストップリストの利用

Companion アルゴリズムなどでは，結果に与える影響が極端に大きいハイパーリンクを集めたストップリストを作成し，評価対象から除外している．

この方法は処理は簡単だが，実際のシステムにおいては，処理結果を見ながらストップリストを管理する作業が必要になる．

4·3 Web ページのグループ化

既存のサーチエンジンの情報閲覧が困難である原因の一つは，検索結果をファイル単位でしか扱えないことである．そこで，Web ページをグループ化する研究が行われている．これは，文書の内容の類似性に基づくクラスタリングと，ハイパーリンク接続解析に基づく Web 情報空間内のコミュニティの発見，そして Web 文書の構造推定に大きく分類できる．

§1 内容の類似性に基づくクラスタリング

自然言語処理の分野で盛んに行われてきた研究であり，例えば Web ページに適用した例として Zamir らの研究が存在する [Zamir 99].

§2 Web 情報空間におけるコミュニティの発見

例えば，ある Web ページをもとにオーソリティとハブを求めることで，関連の深い Web ページを発見する方法が提案されている [Dean 99, 豊田 00].

同様に，われわれは，Web ディレクトリの各カテゴリに含まれる複数の Web ページをシードセットとして関連の深い Web ページを発見することで，Web ディレクトリの保守を省力化する研究を行っている [原田 01].

また，ハブとオーソリティからなる完全二部グラフを，何らかのトピックに関するコミュニティと定義することで，大規模なハイパーリンクのグラフ構造のなかから多数のコミュニティを発見する方法も提案されている [Kumar 99, 村田 00].

§3 Web 文書の構造推定

この方法は，サーチエンジンの検索結果を見やすくするために使用されている．永藤らによる研究では，ページ群の候補としてディレクトリ単位に分割した後に，文書構造を推定するためのリンク情報と，内容の類似性を推定するための単語の出現頻度をもとにした特徴量ベクトルを用いて再帰的にページ群を求めている [永藤 98]. ただし，同じ文書の異なるページは別の事柄を記述するためにある場合が多く，必ずしも内容が類似しているとはかぎらないので，グループ化率が低い．

われわれはページ群の候補としてディレクトリ単位に分割した後に，サーバ内部とサーバ外からのリンク情報と，URL の命名に対するいくつかの経験則をもとにインデックスページを抽出し，同一のインデックスページを持つページ群を統合することでグループ化を行っている [風間 00b]. この手法で生成されるページ群は比較的小さく，同一サイト内の複数の文書を表示するのに適している．

高野らは，逆に同一ホスト内のページをページ群の候補とし，被参照リンク数を元に分割してグループを求めている [高野 00]. この手法で生成されるページ群は比較的大きく，サイト単位の表示に適している．

ほかに，小林らはタグの種類と順序のような Web ページのスタイルを用いている [小林 99]. ほかにも META 要素，背景色などの多くの情報が利用できると思われる．

5. 検索履歴

サーチエンジンの検索履歴を用いて多数の利用者の行動を分析し，検索支援に利用する研究も行われている．

われわれはサーチエンジンが検索結果を表示する際に，その検索結果に関連する語を提示することで，絞り込み検索を支援する方法を提案した [原田 97]. この方法では，まずサーチエンジンの利用履歴から，質問に用いられた語と，その検索結果のなかから閲覧された Web ページの対応を抽出して，表に記録しておく．そして，検索時にはこの表を参照して，検索結果中の Web ページのうち，かつて閲覧先として選ばれた Web ページを列挙し，それらの多くと対応している語を関連語として提示する．実際に他の利用者が検索に用いた語が提示されるために，類義語辞書を用いる方法より，利用者にとって参考になりやすい．

また，川前らはサーチエンジンの検索履歴から，利用者が質問中の検索語を詳細化していくシーケンスを抽出し，それらから検索語どうしの関係を木構造の関連図として整理する方法を述べている [川前 00]. この関連図に含まれる語で検索を行った利用者は，ほかの利用者がどのように質問を詳細化していったかを追体験できる．

大久保らは，サーチエンジンで検索された語の頻度から，利用者全体の情報需要のトレンドを分析する方法を示した [大久保 98]. この方法では，同一利用者が短い時間内に使用した検索語は同じ目的で利用されているという仮定に基づいて，例えば「桜」と「花見」のような単語をグループ化して扱うことで，類義語が別々に集計されないように工夫している．

Glance は，検索結果集合の重複をもとに検索質問どうしを関連づけることで，他の利用者が用いた二つ以上の単語からなる質問を提示する方法を提案している [Glance 01].

6. おわりに

Web 情報検索の研究は，最初に述べた課題のために非常に困難であると誤解している研究者も多いと思われるが，これを解く鍵は，情報の内容だけでなく，ネットワーク社会における情報間の関係や評価を観測して，利用者の支援に利用することにある．そこで，本稿では，それを裏付けるさまざまな研究や実例を紹介した．

現在の Web 情報検索は必ずしも使いやすいものでなく，Gnutella や Pinot のような新しいアーキテクチャも含めて，研究の余地は多い．今後，この分野へよりいっそうの研究者の参加を期待したい．

◇ 参 考 文 献 ◇

[赤峯 00] 赤峯 享, 河合英紀, 小西弘一, 菊池賢太郎, 福島俊一: PCクラスタを用いた並列全文検索エンジン（1）—概要—, 情報処理学会第60回全国大会, 2000.

[Albert 99] Albert, R., Jeong, H., Barábasi, A.L.: Diameter of the World Wide Web, Nature, No. 401, pp. 130-131, 1999.

[Bharat 98b] Bharat, K., Henzinger, M. R.: Improved Algorithms for Topic Distillation in Hyperlinked Environments, Proc. of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[Bharat 98a] Bharat, K., Broder, A.: Estimating the Relative Size and Overlap of Public Web Search Engines, Proc. of 7th World Wide Web Conference, 1998.

[Broder 00] Broder, A., et al.: Graph structure in the web, Proc. of 9th World Wide Web Conf., 2000.

[Bowman 94] Bowman, C. M., et al.: The Harvest Information Discovery and Access System, Proc. of 2nd World Wide Web Conference, 1994.

[Brin 98] Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proc. of 7th World Wide Web Conference, 1998.

[Chakrabarti 98] Chakrabarti, S., et al.: Automatic resource compilation by analyzing hyperlink structure and associated text, Proc. of 7th World Wide Web Conference, pp. 65-74, 1998.

[Dean 99] Dean, J., Henzinger, M. R.: Finding Related Pages in the World Wide Web, Proc. of 8th World Wide Web Conference, 1999.

[Francis 95] Francis, P., Kambayashi, T., Sato, S., Shimizu, S.: "Ingrid: A Self-Configuring Information Navigation Infrastructure," Proc. of 4th World Wide Web Conference, 1995.

[Glance 01] Glance, S. G. : Community Search Assistant, IUI' 01, pp. 91-96, 2001.

[Gnutella] Gnutella: http://www.gnutella.wego.com/

[Goldberg 92] Goldberg, D., et al.: Using collaborative filtering to weave an information tapestry, Communications of the ACM, Vol. 35, pp. 61-70, 1992.

[Google] Google: http://www.google.com/

[Gravano 99] Gravano, L., Garcia-Molina, H., Tomasic, A.: "GlOSS: Text-Source Discovery over the Internet," ACM Transactions on Database Systems, 1999.

[原田 00] 原田昌紀, 風間一洋, 佐藤進也: Unicodeを用いたN-gram索引の一実現方式とその評価, 情報処理学会研究報告, 2000-NL-136-17, 2000.

[原田 01] 原田昌紀, 風間一洋, 佐藤進也: 参照共起分析のWebディレクトリへの適用,情報処理学会研究報告, 2001-FI-61-7, 2001.

[InfraSearch] InfraSearch: http://www.infrasearch.com/

[InfoNavigator] InfoNavigator: Inside of InfoNavigator, http://infonavi.infoweb.ne.jp/ap/

[JXTA] Project JXTA: http://www.jxta.org/

[Kleinberg 99] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, the Journal of the ACM, 1999.

[Lawrence 98] Lawrence, S., Giles, C.L.: Searching the World Wide Web, Science, No. 280, pp. 98-100, 1998.

[Lawrence 99] Lawrence, S., Giles, C.L.: Accessibility of information on the web, Nature, No. 400, pp. 107-109, 1999.

[村田 00] 村田剛志: Webにおけるコミュニティの発見, 第47回人工知能学会知識ベースシステム研究会資料, SIG-KBS-9904, pp. 79-84, 2000.

[NOW] The Berkeley NOW Project: http://now.cs.berkeley.edu/

[永藤 98] 永藤拓宏, 遠山元道: ページ群への分割を利用したWWW検索エンジン, 第9回データ工学ワークショップ, 1998.

[野田 97] 野田良平: AltaVista Searchについて, DECUS News, 1997.

[ODIN] ODIN: http://odin.ingrid.org/

[大久保 98] 大久保雅且 ほか: WWW検索ログに基づく情報ニーズの抽出, 情報処理学会論文誌, Vol. 39, No. 7, pp. 2250-2258, 1998.

[RedHat 00] RedHat: RedHat Linux is an Essential Component of Google's Innovative Web Searching Technology, Press Release, May 30, 2000.

[Selberg 95] Selberg, E., Etzioni, O.: Multi-Service Search and Comparison Using the MetaCrawler, Proc. of 4th World Wide Web Conference, 1995.

[SEW] Search Engine Watch: http://www.searchenginewatch.com/

[Silverstein 98] Silverstein, C., Henzinger, M., Marais, J., Moricz, M.: Analysis of a very large AltaVista query log, Technical Report 1998-014, Compaq Systems Research Center, 1998.

[Sugiura 00] Sugiura, A., Etzioni, O.: Query Routing for Web Search Engines: Architecture and Experiments, Proc. of 9th World Wide Web Conference, 2000.

[佐藤 99] 佐藤進也, 風間一洋, 原田昌紀: 分散検索アルゴリズムの実装と評価, WIT' 99, 1999.

[豊田 00] 豊田正史: WWWにおける関連コミュニティ群の発見, 情報処理学会研究報告, 2000-DBS-122-40, pp. 307-314, 2000.

[高野 00] 高野 元, 久保進也: サイテーション・エンジン: リンク解析を用いたWWW検索ランキングシステム, データベースシステム 120-2, pp. 9-16, 2000.

[上原 00] 上原 稔, 森 秀樹: 最新情報の検索に適した協調サーチエンジン, bit, Vol. 33, No. 1, pp. 43-48, 2000.

[Kumar 99] Kumar, R., et al. : Trawling the Web for emerging cyber-communities, Proc. of 8th World Wide Web Conference, 1999.

[風間 00a] 風間一洋, 原田昌紀, 佐藤進也: ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法, 情報処理学会研究報告, 2000-FI-59-3, 2000.

[風間 00b] 風間一洋, 原田昌紀, 佐藤進也: サーチエンジンの検索結果のマルチレベルグルーピングの評価, コンピュータソフトウェア, Vol. 17, No. 4, pp. 58-69, 2000.

[川前 00] 川前徳章, 青木輝勝, 安田浩: ユーザ履歴を活用した検索システム, 情報処理学会研究報告, DB-122-15, pp. 113-120, 2000.

[小林 99] 小林伸行, 北川文夫: WWW上のページセットの抽出とそれを用いた検索, 第10回データ工学ワークショップ, 1999 .

[原田 97] 原田昌紀, 清水 奨: WWW検索システムにおける不特定多数の操作履歴の活用, 情報処理学会研究会報告, 97-DPS-81-11, 1997.

[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.

[山名 00] 山名早人: 検索エンジンと高速ページ収集技術—分散型WWWロボット実験, bit, Vol. 32, No. 12, pp. 72-79, 2000.

[Zamir 99] Zamir, O., Etzioni, O.: Grouper: A Dynamic Clustering Interface to Web Search Results, Proc. of 8th World Wide Web Conference, 1999.

2001年5月13日 受理

著 者 紹 介

風間 一洋

1988年京都大学大学院工学研究科精密工学専攻修士課程修了. 同年日本電信電話（株）入社. 現在NTT未来ねっと研究所主任研究員. 分散協調処理, 情報検索の研究に従事. 日本ソフトウェア科学会, ACM各会員.

原田 昌紀

1974年生. 1998年東京大学大学院総合文化研究科広域科学専攻修士課程修了. 同年日本電信電話（株）入社. 情報検索の研究に従事. 現在NTT未来ねっと研究所所属. 情報処理学会会員.