データベースシステム 109－44
（1996. 7. 26）

WWWデータ資源検索システムの実装と評価*

西村 英樹† 河野 浩之‡ 長谷川 利治‡

†シャープ株式会社 ‡京都大学大学院工学研究科応用システム科学専攻

データマイニング技術を応用した検索絞り込み機能や，クライアントからのネットワークコストを含む評価機能などをもつWWWデータ検索システムの開発を行っている．本稿では，WWWサーバからデータを収集するロボットプログラムと，そのデータの格納を行うデータベースの実装に関して述べる．次に，検索システムへのアクセスログを元に，問合わせやアクセス傾向などの分析を行う．さらに，重み付き相関ルール導出アルゴリズムによって導出されたキーワードの絞り込み性能の評価を行い，提示されたキーワードがユーザの検索要求に対してどの程度有効かに関して詳しく調べる．

キーワード：WWW資源検索システム，ロボットプログラム，重み付き相関ルール，データマイニング，データベースからの知識発見

Implementation and Evaluation of WWW Search System RCAAU

Hideki NISHIMURA† Hiroyuka KAWANO‡ Toshiharu HASEGAWA‡

†Sharp Corporation
‡Department of Applied Systems Science, Faculty of Engineering, Kyoto University

We have been developing WWW search system with the several advanced functions, such as keywords focusing by a data mining techniques and network characteristics evaluation. In this paper, we explain implementation of WWW robot which collects data from Web servers and stores several attributes into database. Then, based on http access log, we analyze keywords of queries and its embedded tendency. Moreover, in order to ivestigate thoroughly how effective our functions are for users, we evaluate the quality of keywords derived by weighted association rule.

Keywords: WWW Search Engine, Robot Program, Weighted Association Rule, Data Mining, Knowledge Discovery in Databases

*連絡先: 〒606 京都市左京区吉田本町 京都大学大学院工学研究科応用システム科学専攻 河野 浩之
Tel: (075) 753-5513, Fax: (075)761-2437, E-mail: kawano@kuamp.kyoto-u.ac.jp

1 はじめに

現在，増大するWWW(World Wide Web)の資源に対して多様な検索システムが登場している．しかし，広い概念をもつキーワードを用いた検索においては，多数のドキュメントが条件を満たし，質の高いドキュメントを素早く閲覧することを困難としている．

そこで，我々は，データマイニング(Data Mining)に関連する技術[3, 8, 1]を応用し，WWWデータ検索システム(RCAAU)の開発を行っている．これまでに，キーワードに重みを与えることによって，相関ルール導出アルゴリズム[7]を，重み付き相関ルール導出アルゴリズムとして拡張し，検索集合において関連度の高いキーワードをルール（知識）として導出する検索システムを実装した[6]．さらに，クライアント側のブラウザからのネットワークコスト評価機能を含むデータベースブラウザとしてJava appletの実装を行った[2]．また，関連するキーワードをルールとしてユーザに提示する検索システムの運用実験を行っている[1]．

そこで，本稿では，検索システムの運用実験によって収集されたアクセスログを元に，アクセス傾向と問合せなどの分析を行う．特に，重み付き相関ルールによって導出されたキーワードを用いた絞り込み過程を評価することによって，ユーザの検索要求に対して関連キーワードを提供することの有用性に関して詳しく述べる．

第2章で，開発中の検索システムの概要を簡単に述べ，第3章では，各WWWサーバの負荷を考慮してドキュメントをバランス良く取得する探索アルゴリズムと，関連するデータを格納するデータベースの構造について述べる．次に，第4章で，アクセス傾向と問合せなどの分析，キーワード提示による検索過程の状況に関して調べる．第5章では，今後の検索システムに必要とされる事項について分析結果を元にして考察し，第6章に結論をまとめる．

2 検索システム – RCAAU –

図1に示すように，検索システム(RCAAU)は，(1)WWWドキュメント収集エージェント，(2)検索要求に対してリアルタイムでルール導出を行うサーバ，から構成されている．

[1]検索システムへのアクセスは，http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/から可能としている．検索結果のコメントにおいて，データベースブラウザにおいて表示すべきデータを含めた提供を行っている．

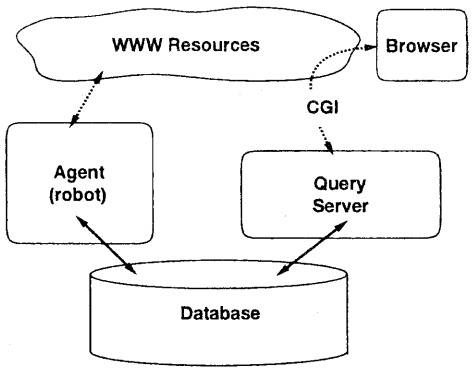

図1: RCAAUの構造

WWWドキュメント収集エージェント

ロボットプログラムのガイドライン[5]に準拠して実装されているエージェント(agent)が，ネットワークからのWWWドキュメント収集を行う．エージェントは，WWWドキュメントを収集・解析・圧縮し，データベース(database)へと格納する．また，データベースのデータを解析しながら，WWWデータ空間の渡り歩きを行う．ただし，ネットワークへの負荷を最小限に抑えるため，現在は単一プロセスで実装している．なお，第3章において，データベース構造とWWWサーバの探索アルゴリズムについて詳しく述べる．

検索処理サーバ

検索処理サーバ(query server)は，ユーザからの検索要求に対して応答する．まず，WWWブラウザ(browser)によって検索記述が要求されると，CGI(Common Gate Interface)を通じて検索処理サーバへ通知される．次に，データベースから条件を満すデータ集合を求め，それらのデータ集合に対してデータマイニング操作が行われる[6]．検索結果は，HTML形式によってユーザへと提示される．

3 検索システムの実装

3.1 データ収集エージェントの探索アルゴリズム

我々は，WWWデータ資源の探索において，

- ネットワークに高負荷をかけないドキュメント収集

- 重要なドキュメントを優先してバランス良く収集

の二点を中心に考慮して，アルゴリズム 1,2の設計を行った．なお，データベースを解析することによって，ハイパーテキストのリンク構造を用いた探索を実現しており，かつ，複数ロボットの知的協調動作をも目標としていることから，本稿ではエージェントと呼ぶことにする．

アルゴリズム 1 【エージェントによる探索】

1. サーバのリスト *(slist)* を作成．
2. サーバ毎にドキュメントのツリー *(dtree(s))* を作成．
3. *slist* より探索すべきサーバ s を選択．
4. 選択されたサーバ s 上の未探索ドキュメント $d \in dtree(s)$ を *1* つ選択．
5. d をネットワークを通じて取得．
6. d を *dtree(s)* から除去．
7. d を解析し，新しいサーバ・未取得のドキュメントを *slist*, *dtree* に追加．
8. 未取得のドキュメントが無ければ終了．(もしくは，ドキュメント変化の有無を再探索する．)
9. ステップ *3*へ． □

上記アルゴリズム中のサーバ選択は，サーバリストの先頭から順番に選択することとした．しかし，選択されたサーバ s における未取得ドキュメントは，一般に複数存在する．そこで，被参照回数の多いドキュメントが重要であると考え，各ドキュメント d に対して，「d を参照しているドキュメント数」により重要度 $v(d)$ を定義し，$v(d)$ を用いてドキュメント選択手順を実装した[2]．

アルゴリズム 2 【ドキュメント選択】

1. サーバ s 上の未取得ドキュメント $\forall d \in dtree(s)$ に対して全ての $v(d)$ を計算．
2. 最大の $v(d)$ をもつ d を選択． □

なお，図2に，実装アルゴリズムにより取得される順序の一例を簡単に示しておく．図では，被リンク総数を考慮した重要度の性質によって，深い位置にあるドキュメント 5 が，浅い位置にあるドキュメント 7 よりも先に取得されている．

[2]WWW データは，ドキュメントの質のばらつきが大きく，複数サーバでのリンクの一貫性の保持が困難であり，統一的なデータモデルを見い出すことが困難な構造を形成している [4].

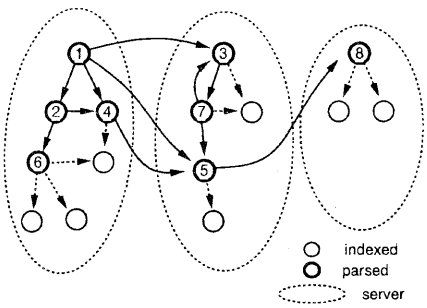

図 2: 探索アルゴリズム

3.2 データベース構造

本節では，実時間で効率良く，検索並びに計算コストの高いデータマイニング処理を実行するために必要となるデータベース・スキーマに関して述べる．

・Keyword テーブル

$$K_m : (URL_{m1}, W_{m1}), \ldots, (URL_{mn}, W_{mn})$$

キーワード K_m に対して，ドキュメント URL_{mn} と重み W_{mn} の組のリストが格納される．

・URL テーブル

$$URL_m : Attr_m, (K_{m1}, W_{m1}), \ldots, (K_{mn}, W_{mn})$$

URL_m に対する，タイトルや作成日などからなる $Attr_m$ と，ドキュメントが含むキーワードと重みの組 (K_{mn}, W_{mn}) が格納される．

・Link テーブル

$$URL_m : URL_{m1}, URL_{m2}, \ldots, URL_{mn}$$

URL_m が URL_{mn} を参照するハイパーテキストのリンク構造を格納し，WWW 探索順序の決定に用いる．

なお，データマイニング処理のコストが大きいため，検索条件を多数の要素が満す場合，Keyword テーブルの重み W_{mn} を用いて URL 集合の要素数を制限し，URL テーブルからデータマイニングの実行対象となる集合を生成する．

3.3 収集されたデータの性質

1996 年 6 月 1 日現在のデータベースの状況を表1に示す．データベースのサイズには，インデックスなどのデータも含まれていることに注意すると[3]，

[3]インデックスの実装に GNU DBM を使用している．

ドキュメントの属性を圧縮してデータベースに格納していることから，ハードディスク占有量は比較的小さく抑えられていると考えられる．

表 1: データベースの状況 (1996 年 6 月 1 日現在)

登録キーワード数	237,673 語
検索可能 URL 数	743,484 URL
うち国内 (%)	619,721 URL(83%)
収集 URL 数	250,352 URL
収集サーバ数	9,012 サーバ
未収集 URL 数	369,369 URL
ハードディスク占有量	433 MB
1URL あたり占有量	610 B/URL

次に，1996 年 2 月 28 日 から 同 6 月 1 日 までの収集した URL 数 (parsed) と検索可能な URL 数 (indexed) の推移を，図 3に示す．

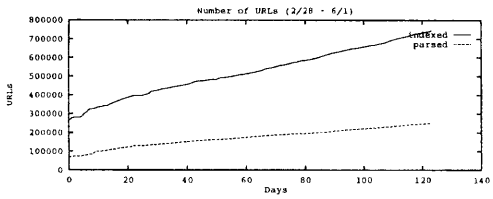

図 3: URL 数の推移

取得ドキュメントが増加しても，新たな URL が含まれているため，両者はほぼ一定の比率（約 3 程度）を示していることが，図 3から分かる．

4 ユーザの検索要求と問合せ記述

本章では，1996 年 4 月中のキーワード検索システムへのアクセス 51,105 件に関して詳細な評価を行う．

4.1 アクセス件数の変化

日別・時間別のアクセス件数を，それぞれ，図 4，図 5に示す．なお，図 4によれば，アクセス数の増減は一週間周期であり，休日にアクセスが少なくなること，特に co ドメインにおいて，この傾向が強いことが伺える．

また，図 5からは，アクセス数の時間帯による増減は，ドメインにより微妙にずれており，co ドメインに比べ，ac ドメインは $\frac{\pi}{3}$，or ドメインは $\frac{2\pi}{3}$ 程度の位相遅れがある．つまり，時間帯により，各ドメインからのアクセス数比率が，一定の傾向をもって変化していると言える．

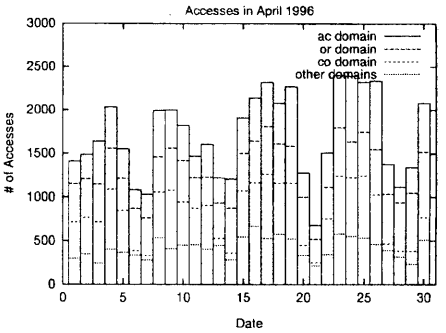

図 4: 日別アクセス数の変化

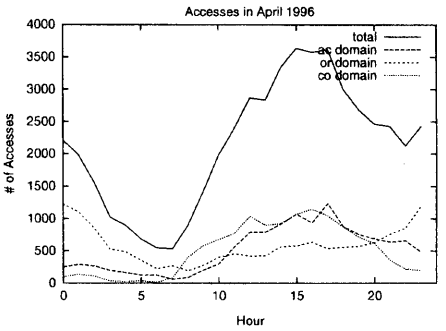

図 5: 時間別アクセス数の変化

表 2: アクセス内訳，括弧内は割合 (%)

ドメイン		ac	or	co	その他	計
20 件表示	完全	8,921 (17.5)	9,648 (18.9)	8,449 (16.5)	8,878 (17.4)	35,896 (70.2)
	前方	614 (1.2)	809 (1.6)	606 (1.2)	532 (1.0)	2,561 (5.0)
50 件表示	完全	703 (1.4)	700 (1.4)	646 (1.3)	677 (1.3)	2,726 (5.3)
	前方	92 (0.2)	182 (0.4)	70 (0.1)	78 (0.2)	422 (0.8)
100 件表示	完全	488 (1.0)	591 (1.2)	736 (1.4)	723 (1.4)	2,538 (5.0)
	前方	15 (0.0)	53 (0.1)	78 (0.2)	91 (0.2)	237 (0.5)
200 件表示	完全	1,423 (2.8)	1,826 (3.6)	1,293 (2.5)	1,388 (2.7)	5,930 (11.6)
	前方	183 (0.4)	267 (0.5)	166 (0.3)	179 (0.4)	795 (1.6)
計	完全	11,535 (22.5)	12,765 (25.0)	11,124 (21.8)	11,666 (22.8)	47,090 (92.1)
	前方	904 (1.8)	1,311 (2.6)	920 (1.8)	880 (1.7)	4,015 (7.9)
合計		12,439 (24.3)	14,076 (27.6)	12,044 (23.6)	12,546 (24.5)	51,105 (100)

4.2 アクセス件数の内訳

アクセス件数のオプション選択に関する内訳を，表2に示す．各ドメインでの差は殆んど認められず，検索オプションとして用意した「20件，50件，100件，200件」，「完全一致，前方一致」では，デフォルト値として設定した「20件表示，完全一致」が最も多く，全体の70%を占める．また，20件以外や，前方一致に変更する場合も，それぞれ，25%，8%程度存在している．

ここで，20件についで多い指定が，最多の設定である「200件」であり，表示件数の選択において二極分化していることが分かる．

4.3 絞り込み機能の利用状況

次に，検索要求を出したユーザが，提示した関連キーワードを用いて絞り込みを行うかを調べ，表3に示す．

表 3: 絞り込み回数比

項目	回数	回数比 (%)		
総検索数 (含絞り込み)	51,105	100	–	–
入力フォームからの検索	35,305	69.1	100	–
絞り込みによる検索	15,800	30.9	–	–
1回目の絞り込み	13,467	–	38.1	100
1回目のみ絞り込むもの	11,263	–	31.9	83.6
複数回絞り込むもの	2,204	–	6.2	16.4
キーワードを追加入力	1,381	–	3.9	–
絞り込まないもの	20,457	–	57.9	–
否定キーワードを指定	629	–	1.7	–

検索要求者のうちの38.1%は，導出された関連キーワードを用いて絞り込みを行っており，複数回絞り込む要求も全体の6.2%を占める．また，入力フォームでキーワードを直接追加して検索する者は3.9%に留まっていることから，絞り込みキーワードの提示は有効に利用されていると言える．また，絞り込みを行っていない57.9%には，検索条件を満す件数が少ない場合などシステムから付与される順序付けで十分な場合が多く含まれていると考えられる．

なお，否定条件にキーワードを指定するものは，フォームから入力する者のうちのわずか1.7%であることから，適切なNOT条件を検索式を記述することは非常に困難であることが推測される．

4.4 導出された関連キーワードの被覆性

本節では，導出された関連キーワードの被覆性について評価する．ここでは，1996年4月のアクセス回数の多い20キーワードを用いて，導出された関連キーワードの被覆特性を表4にまとめた．但し，一段階の絞り込みのみ対象とした．

表 4: 関連キーワードの被覆特性

調査対象キーワード	アクセス回数上位20語
平均ヒット件数	714
平均導出キーワード数	8.6
平均絞り込み率	8.8 %
平均絞り込み不可割合	51.9 %

一回の絞り込みで平均8.8%にまで絞り込まれており，我々の実装は非常に収束の早い絞り込みを与えていることがわかる．例えば，714件の検索集合において成立する特徴を把握し，目的のドキュメントを探し出すコストは非常に大きいと考えられるが，8.8%に相当する62件に絞り込めば，検索集合の特徴をブラウジングすることによって把握することもかなり容易になると思われる，また，必要に応じて，絞り込みを繰り返せば，さらに5件程度へと絞り込むことが可能である．

しかしながら，どの導出されたキーワードを用いて絞り込んでも到達不可能なドキュメントが51.9%あることが明らかとなっている．そのため，検索キーワードに対する集合の大きさに応じて閾値を動的に変化させるなどの改善が必要と考えられる．ただし，上記評価では行っていないが，文献[4]に述べたように，リンク構造による被覆領域が考慮できることに注意しなければならない．

また，平均導出キーワード数が8.6語となっているのは，実時間性を重視しているために提示キーワード総数の上限を10語としているからである．これは，導出アルゴリズムの実行環境を並列化するなどして，データマイニング処理速度を向上させることによって導出キーワード数を増加すれば，被覆範囲の部分的な改善も可能である．

ただし，被覆範囲を大きくするために，シソーラスを用いるなどすることは必ずしも良い結果を与えないと考えている[4]．なぜなら，多数の独立したユーザが異なる方針でドキュメント作成を行うため，質の揃った属性（タグ）・キーワードを求めることが非常に困難であるからである．さらに，自由なリンク構造を許しているため，関連するドキュメントのクラスター構成が非常に緩やかになっており，適切な関連キーワードのベクトルを構成することも難しくしている．

5 主要項目に対する考察

本章では，前章までの議論をもとに，より良い効率的な検索結果を提供するために必要となる点に関して考察する．

探索アルゴリズムの実装

全てのサーバは同等ではなく，頻繁にWWWデータが修正されるサーバや負荷が恒常的に高いサーバも存在する．そこで，参照されるリンク数以外に，各サーバの性質に応じた重みの利用が考えられる．

データベースの実装

多量のハイパーテキストデータを解析・圧縮することによって，比較的小さく格納する実装となった．しかし，電子化ドキュメント数は急増傾向にあり，種々の特性をもつ各種検索システムをネットワーク上に分散協調させる研究 [4] が必要になる．

検索要求への対応

検索要求の変化を分析し，時間帯やドメインによる検索傾向などに応じて，ドキュメントの重みなどを用いて検索結果の精度を変化させることが必要と思われる．また，デフォルトのオプション利用が大多数を占めるが，最多件数の検索結果表示を要求する場合も多いことから，ログに記録された頻出キーワードに応じて検索結果にバイアスを与える仕組みの有効性に関する考察も必要である．

導出されたルール（関連キーワード）の品質

関連キーワードを提示する機能は，目的のリソース発見に必要となる検索記述を改善するが，検索条件によっては，導出コストが高くなるために，その機能が十分に発揮できないことがある．アルゴリズムの閾値や補集合を与えることによって被覆性を高めるだけではなく，データマイニングの対象をハイパーテキスト特有のリンク構造へと広げ，リンクに関するルールを抽出すること [4] も併せて考える必要がある．

6 むすび

本稿では，サーバの負荷やドキュメントの重要度を考慮した探索により，各WWWサーバからバランス良く低負荷にてドキュメントを収集し，多量のハイパーテキストデータを適切に格納するデータベースの実装について述べた．また，ユーザからの検索要求のログにおける傾向を元に，アクセス件数が，ドメインによって異なった傾向の時間変化を示すことや，デフォルトのオプション選択で検索する場合が多いが，最多件数の表示を要求する場合が次いで多いことなどを示した．

なお，絞り込み支援機能は目的のリソースを発見することに非常に強力である一方，その機能が有効に働き難い状況が存在する場合があることがわかった．今後，検索要求に関するログを元に，頻出キーワードに対して精度の高いルールをあらかじめ導出してキャッシュすること，導出するルールの精度をリンク構造を考慮して向上させることにより，より質の高い検索を可能とするルール導出などの研究が必要と考えられる．

謝辞

本稿の一部は文部省科学研究費重点領域「分散発展型データベースシステム技術の研究 (08244103)」のもとでの研究成果による．また，本稿は，筆頭著者の京都大学在学中の成果を元に，新たに記録されたログなどを用いてまとめたものとなっている．

参考文献

[1] U. Dayal, P. M. D. Gray and S. Nishio Eds.: "Proceedings of the 21st International Conference on Very Large Data Bases," Zurich, Switzerland (1995).

[2] 伊藤耕一郎, 西村英樹, 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をもつ検索インタフェースのWWWデータへの適用," 電子情報通信学会 1996 年総合大会, Vol.D, pp.307–308 (1996).

[3] 河野 浩之, 西尾 章治郎, Han, J.："データベースからの知識獲得技術," 人工知能学会誌, Vol.10, No.1, pp.38–44 (1995).

[4] 河野浩之, 長谷川利治："WWW データ資源検索におけるデータマイニング手法," 情処研報 96-DBS-108, No. 45, pp.33–40 (1996).

[5] M. Koster: "Guidelines for Robot Writers," http://info.webcrawler.com/mak/projects/robots/guidelines.html.

[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), pp.79–84 (1996).

[7] R. Srikant and R. Agrawal: "Mining Generalized Association Rules," Proceedings of the 21st International Conference on Very Large Data Bases (Eds. by U. Dayal, P. M. D. Gray and S. Nishio), Zurich, Switzerland, pp.407–419 (1995).

[8] O. R. Zaiane and J. Han: "Resource and Knowledge Discovery in Global Information Systems: A Preliminary Design and Experiment," Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, pp.331–336 (1995).