筑波大学技術報告
24: 62-65, 2004

筑波大学内 Web データの全文検索システムの構築

佐藤守
筑波大学 研究協力部研究協力課（学術情報処理センター）
〒305-8577 茨城県つくば市天王台 1-1-1

概要

筑波大学内 Web データを対象とした検索システムを構築し，筑波大学公式ページ内を検索対象範囲に限定した検索サービス[1]として利用を開始した．本稿では日本語全文検索システム構築と運用に関する取り組みを紹介しながら，ログ統計の有効活用について報告する．

１．はじめに

筑波大学の公式ページでは，学外の方に対する情報提供を行うサービスとして Web 検索システムを導入し，大学が保有する大量のデータの中から目的の情報を効率よく提供するための手段として活用されている．この検索システムは，検索対象が中規模程度でも検索要求に対する応答時間の高速化が期待できる Namazu[2]を検索エンジンとして採用するとともに，各検索対象サイトからのデータ収集を行うために wget[3]というソフトウェアを利用している．収集したデータを事前にインデックス化する処理を含め，これら一連のデータ更新を無人で実施するためのスクリプトを作成することによって，定期的に自動実行することで円滑な運用が可能となった．（図 1）

２．学内 WEB データの収集方針

2.1 学内サイトの不特定収集による予備評価

検索システム構築当初(2002 年)は収集対象の URL を特定せずに筑波大学公式ページを起点とし，そこからリンクを辿り学内サイトについてデータ収集を行った．起点からの収集深度を 10 に設定して Web データの収集を行った際, 216 サイト(重複分を除く)の情報が集まった．収集対象のデータタイプを指定せず画像等も全て収集していたため，合計ファイル容量は約 13GB になり約 14 時間を費やした．同じ収集を行った場合に，現在では図書館情報大学統合分が追加となり，収集時間とデータ量がさらに増加することになる．また，このようなアクセス間隔を設けずに大量のデータを連続的に収集する方法では，収集対象の各サイトに対して高負荷となるため改善する必要があることが判明した．

2.2 特定サイトのデータ収集による評価

次に収集対象データタイプを HTML 文書等に限定することで収集時間とデータ収集量を軽減し，学内の主要な Web サイト(110 サイト)を収集対象とし，深度を徐々に深くしながらデータ収集を実施した．深度毎のデータ収集の傾向について図 2 に示す．ネットワークや各サイトの状態により多少の変動が深度毎にみられるが，収集時間は収集量にほぼ比例し，平均収集速度は 76KB/s であった．さらに高速化したいところであるが，各 Web サーバの運用妨害とならぬようにするためには，収集時のアクセス間隔等を調整しながら時間を掛けて収集することも必要である．

図１．筑波大学公式ページ内の検索

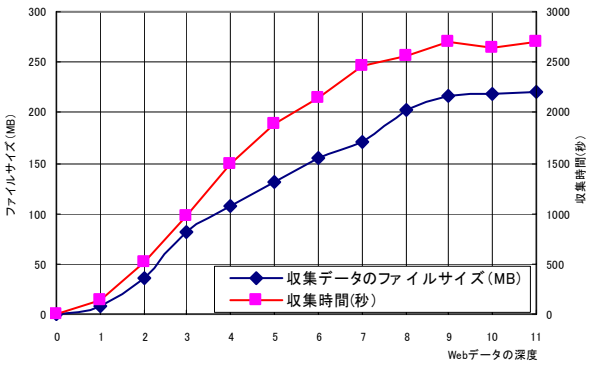

図２．深度毎のデータ収集時間と容量

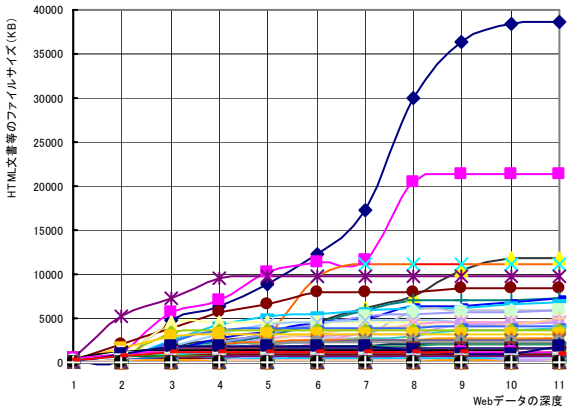

図３．学内 Web サイト別深度毎データ収集量

[1] http://www.tsukuba.ac.jp/~robots/ja/index.html
[2] http://www.namazu.org/
[3] http://www.gnu.org/software/wget/

2.3 最終深度の予測

収集したデータを Web サイト別に分けて深度毎のデータ収集量について調べた．（図 3）
図 2，図 3 ともに深度 10 で飽和状態になることがわかる．そこで，この 2 つの図において深度 11 のデータを深度 inf（無限）でデータ収集した際の数値に置き換えて表示してみた．無限深度での数値が深度 11 での曲線に調和していることから，学内主要 110 サイトにおける収集可能なWeb データの起点（公式ページ）からの最終深度を 11 と予測できる．以上の結果から現在はデータ収集時の上限値として深度 11 に設定し，各 Web サーバへの負荷を軽減するための配慮としてアクセス間隔を 3 秒程度に設定し，30 時間以上かけてゆっくりデータ収集を行っている．

3．検索システムにおける日本語の取扱い

はじめにで述べたように，本検索システムは Namazu を検索エンジンとして採用しているが，その際，日本語文書を検索システム用にインデックス化する過程において単語単位に分割する分かち書き処理に KAKASI[4]を利用している．これには付属の辞書があり単語単位に読みと漢字等のペアで構成（約 12 万件）されている．この辞書だけでも普段使用している一般的な日本語であれば分かち書きの辞書として十分機能する．ただし学内に存在する各種専門分野で使用される専門用語に関する検索では，期待する検索結果を得られない場合もある．検索精度を向上させるためには，分かち書きに用いる辞書の強化を行うことで解決できる．幸いインターネット上には各分野の辞書等[5]が存在し入手可能である．しかし辞書の書式は様々なので KAKASI 用辞書の書式に変換する必要がある．そこでスクリプトを作成し，重複分を排除しながら簡単に変換できるようにした．現在，医学，法学，地球物理学，生命科学の辞書を追加し合計約 17 万件を登録して利用している．必要に応じてさらに辞書の強化を行う予定である．

4．検索システムの利用状況

検索システムを稼動してから現時点で３ヶ月経過した．この間に記録された Web サーバのアクセスログをログ解析ソフトの Analog[6]を用いて，検索システムに関する利用状況を調べた．学内外から平均 28548 件/月の検索があり，日毎の検索件数（図 4）では平均 930 件/日であった．このうち学外からの検索は各月ともに 80%を超えていて全体の平均では 82.6%であった．曜日毎の検索件数（図 5）では平日の火曜日～水曜日が最大となり休日には減少する傾向がわかる．時刻毎の検索件数（図 6）では，16 時～17 時の間でアクセスが最大となることもわかる．これにより検索システムの稼動状況が把握でき，システム作業等を行う場合に，負荷の少ない曜日や時間帯を選択するなど運用面に反映しながら検索応答性能の維持に役立てることができる．

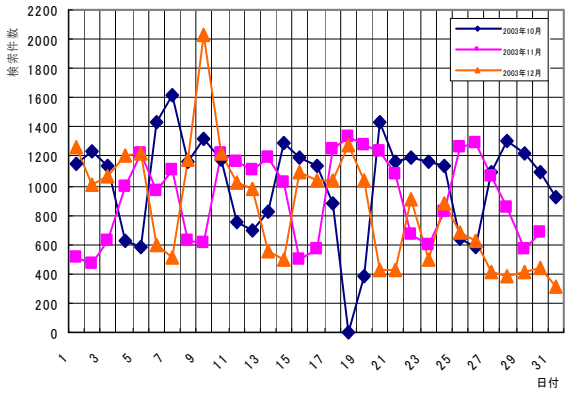

図４．日毎検索件数

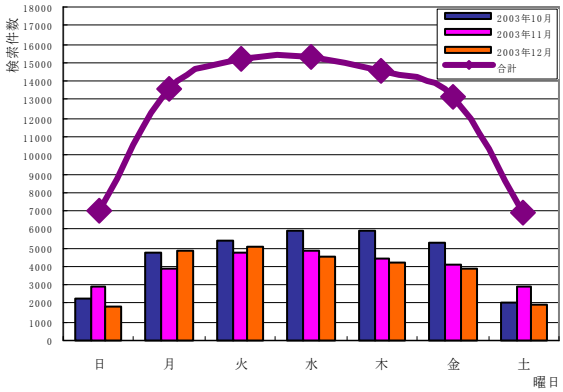

図５．曜日毎検索件数

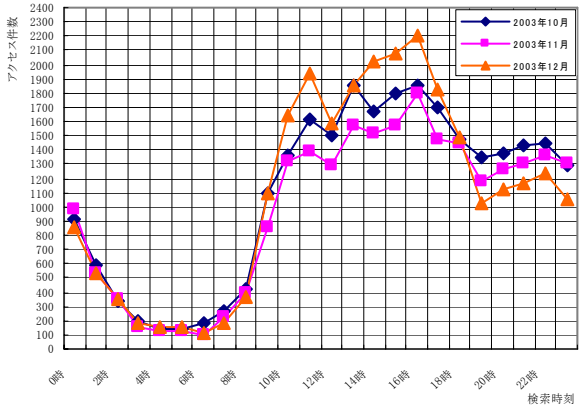

図６．時刻毎検索件数

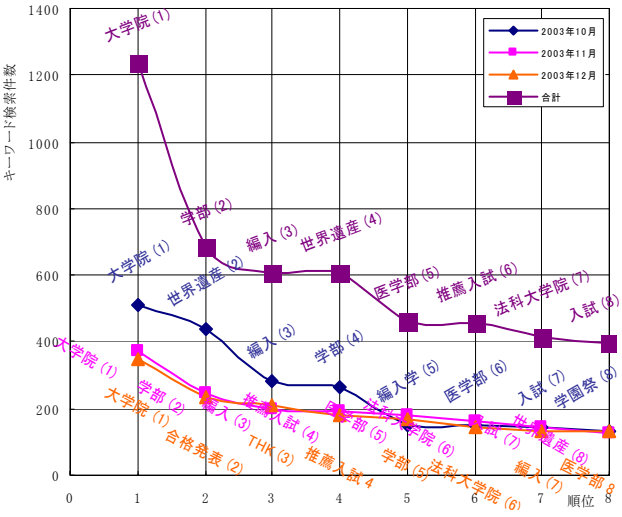

図７．検索キーワードランキング TOP8

[4] http://kakasi.namazu.org/
[5] http://www.kusastro.kyoto-u.ac.jp/~baba/dic/free-dic.html
[6] http://www.analog.cx/

５．検索システムの応用

検索システムでは，検索するキーワードが入力されるため，検索の目的や意図が多少なりとも含まれている．少数のログでは意味を持たないが，大量のログを統計処理して抽出したキーワードの上位ランキングから大学に対する関心事項等が時間経過とともに把握可能である．Namazu のログを解析するツールとして nlview.cgi[7]を利用し，学外からの検索キーワードと検索回数について調べ，上位 8 位までを(図 7)に示した．「大学院」というキーワードが各月とも第 1 位であり学外からの関心が高いことがわかる．また，同じ意味のキーワードでも文字が一文字でも異なると別のキーワードとしてカウントされてしまうため，キーワードの内容とランキングの結果が必ずしも検索の実態を反映しているとは限らない．

６．検索キーワードの総合的な比較

図 7 では「大学院」とは別に「法科大学院」というキーワードもあり，同様に「学部」「医学部」や「編入」「入試」「推薦入試」などから共通のキーワードとして「大学院」「学部」「入」が含まれている．共通キーワードを含む集合について調べることにより総合的な比較ができるようになる．そのためには検索キーワードから指定した文字列に関して解析するツール（図 8）が必要となり次の２つの機能を作成した．

図８．キーワード検索用ツール

6.1 検索キーワードの文字検索機能

これは学内，学外などの対象データ範囲を選択し，指定した期間における検索キーワードについて文字列検索を行い，統計処理してランキング表示する機能である．例えば「大学院」という共通キーワードで学外からの検索について調べてみると，「法科大学院」「連携大学院」「大学院入試」「夜間大学院」「社会人大学院」など 462 種類 2515 件のキーワードがランキング表示される．これにより共通キーワードの母集団同士の比較等が可能になり，総合的な評価を行う際に役立てることができる．

6.2 検索キーワードの文字出現頻度比較機能

これは 6.1 の機能と同様に対象データ範囲と期間を指定し，その間の検索キーワードについて１文字単位に分割し文字の出現頻度を比較する機能である．

[7] http://village.infoweb.ne.jp/~fwnk1502/data/howto2.htm

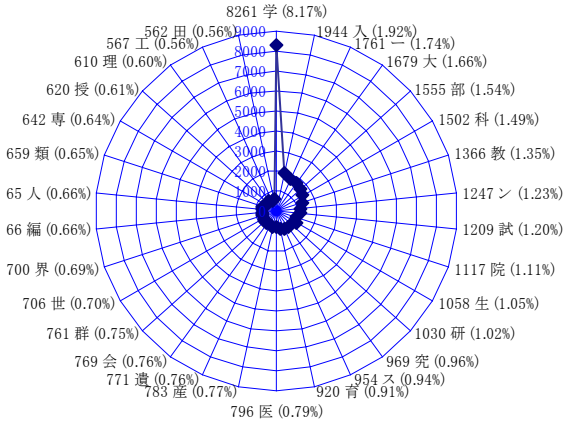

図９．文字出現頻度 TOP30（10 月）

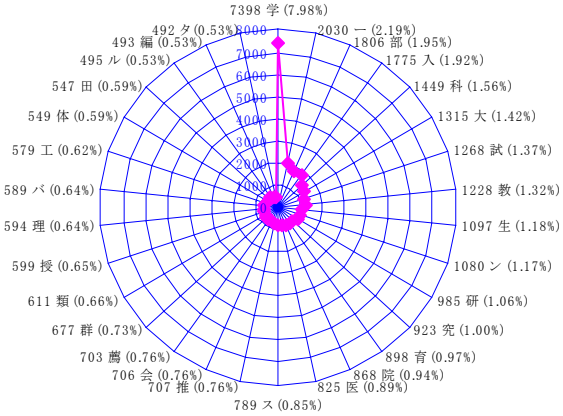

図１０．文字出現頻度 TOP30（11 月）

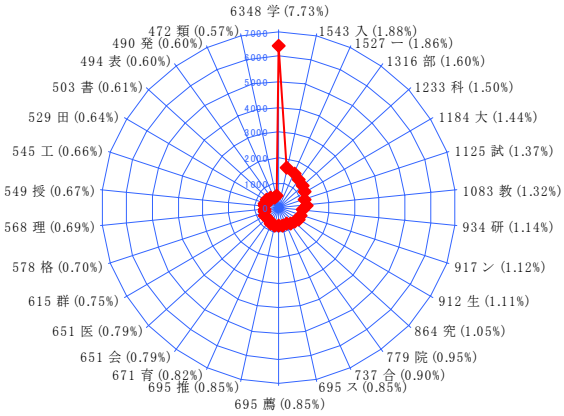

図１１．文字出現頻度 TOP30（12 月）

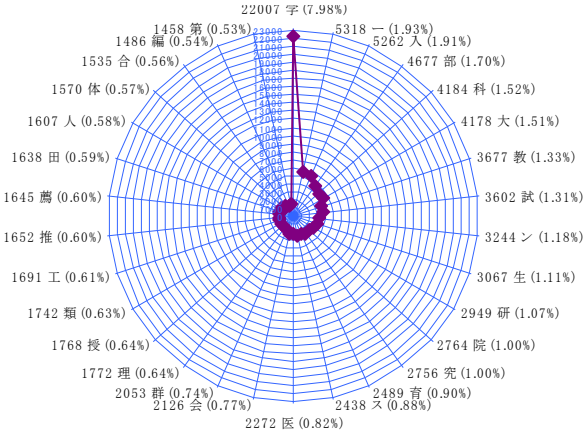

図１２．文字出現頻度 TOP30（合計）

この機能により検索頻度の高い文字を把握でき，さらにその文字を 6.1 の機能で再検索することで，通常のキーワードランキングでは見えないランキングの把握も可能となる．図 9～図 11 は，2003 年 10 月～12 月の各月毎の検索文字出現頻度を表しており，図 12 はその合計である．検索されたキーワードにおける時期的な関心の移り変わりが浮かび上がる．
10 月には「世」「界」「遺」「産」が登場している．
11 月には「バ」「ル」「タ」がランクインしている．「バ」については，「シラバス」や「バスケットボール」等の各月での極端な変動はないのに対し，11 月のバレーは 10 月の約 3.5 倍の増加となり，WorldCup バレーボールの影響によるものと予想できる．「ル」についても同様で，「サークル」等に極端な変動はみられず，通常の「ル」の件数に約 2.3 倍のバレーボールの件数が加算された結果となった．「タ」については，学内の各「センター」等の検索件数に「センター試験」や「アドミッションセンター」等の入試関連の「タ」が増加した結果である．
12 月は，「合」「格」「発」「表」「書」が登場し，「書」は「願書」の増加と「図書館」「証明書」などの件数の合わせ技となった．その他は概ね定番の文字が各月ともに出現している．

6.3 検索キーワードの総合的評価

図 12 を見ると，「学」という文字が出現頻度の第 1 位で他の文字と大差があり，大学として当然の結果となっている．その他の出現した文字と組み合わせると各種検索キーワードが想定できる．第 2 位には，予想外に「ー」という文字が登場した．「サークル」「サッカー」「バレー」「スポーツ」の 4 項目だけで 37.7%を占め，学外からの体育系サークル等に関する関心の高さが見えてくる．「サッカー」と「バレー」の母集団を比較してみると，10 月では「サッカー」が「バレー」の 1.5 倍多かったのに対し，11 月，12 月では「バレー」が 2 倍近くに逆転し，WorldCup 効果が持続していることが観測できる．第 3 位には「入」が登場している．図 7 の合計をみると第 3 位の「編入」609 件と第 6 位の「推薦入試」455 件および第 8 位の「入試」397 件があり，この図では確認できない第 9 位の「編入学」353 件も存在する．これではランキングとして適切に評価することが難しい．そこで 6.1 の機能を利用して「編入」と「入試」について比較してみた．学外からの検索を対象データに選択し，10 月～12 月までの 3 ヶ月間における文字検索を行うと，「入試」を含むキーワードの方が 2811 件（563 種類）となり「編入」の 1469 件（197 種類）より出現頻度が高いことがわかる．このようにキーワードを母集団として比較することで，キーワードに付随する表現の広がりについても把握することができ，単純なランキングでは見えない総合的な比較が可能となる．

7．まとめ

大学独自に検索システムを用意しなくても，Google[8]や Yahoo[9]等の検索エンジンのサイト検索機能を利用すれば，大学の検索機能として活用できることも事実である．その他，商用や独自に開発した検索システムも各種存在し，高機能の検索も可能である．これらのどれを選択するかは，目的や環境に応じて選択すればよい．自前で検索システムを運用する利点は，情報の鮮度を調節できることであり，検索対象を適切に選択することにより，上質な情報を提供できることにある．検索システムを稼動してから現時点で３ヶ月経過したが，学外からの利用率をみるとその重要性を再確認できる．また，検索システムを運用することで，学外からの大学に対する関心や需要などの動向についても時間経過とともに把握することが可能となる．今後は専門分野の辞書を充実させ，全学的に検索対象範囲を拡張することで，検索システム利用者の意図する情報を的確に提供することが可能となり，さらなる利用率の増加が期待できる．これにより副次的な効果も向上し，学外の関心等を把握する手段として有効に活用できる．

検索システムの運用に際し，最後に結論として導く言葉は「アンテナ」である．つまり，検索システムがアンテナ的な役割を果たすことも可能であることを意味する．アンテナの感度を調整しながら得られたキーワードを大学運営の参考としてフィードバックすることにより，大学に求められている役割についての理解を深めることに役立てることも可能である．そのためには，情報の取り扱いに十分注意するとともに，情報の鮮度・質・量を適切に管理しながら検索システムの利便性と信頼性を向上させてゆくことが重要であり，学内における理解と協力が必要となる．

謝辞

検索システム構築の機会を与えて頂き，ご指導とご協力頂いた大学広報課関係各位，学術情報処理センター山口喜教センター長，佐藤聡講師に深く感謝しお礼申し上げます．

参考文献

[1] 馬場肇. Namazu システムの構築と活用，ソフトバンクパブリッシング(2001)
[2] 松木孝幸，高橋基信，太田俊哉．Analog[アナログ]Web アクセスログの高速解析ソフト，九天社(2003)
[3] Namazu2.0 入門
URL:http://www.namazu.org/doc/tutorial.html
[4] KAKASI 漢字→かな（ローマ字）変換プログラム
URL:http://kakasi.namazu.org/
[5] Analog
URL:http://www.analog.cx/
[6] nlview.cgi
URL:http://village.infoweb.ne.jp/~fwnk1502/data/howto2.htm
[7] 法律用語電子化辞書 LKKS
URL:http://icrouton.as.wakwak.ne.jp/pub/kks/index.html
[8] 地球物理辞書
URL:http://www.chibutsu.org/jisho/
[9] 医学用語辞書
URL:ftp://ftp.kusastro.kyoto-u.ac.jp/pub/baba/dic/
[10] ライフサイエンス辞書プロジェクト
URL:http://lsd.pharm.kyoto-u.ac.jp/index-J.html

[8] http://www.google.co.jp/
[9] http://www.yahoo.co.jp/