文献調査支援のためのスタンドアロン型 全文検索システムの構築

渡上将治[1] 村川猛彦[2] 宇都宮啓吾[3] 中川優[2]
[1]和歌山大学 大学院システム工学研究科 [2]和歌山大学 システム工学部 [3]大阪大谷大学 文学部

インターネットに接続する環境のない場所でも利用でき，個人の持つテキストデータを瞬時に検索できる全文検索システムを構築してきた．本稿では，文献調査をシステム側から支援するための機構を構築したので報告する．利用に望まれる性質として，携帯性，正確性，瞬時性を定義し，本システムはすべてを満たすことを示した．さらに再利用性と導入容易性を定義し，検索インタフェースの改良と独自のブックマーク機能の構築，また利用者がファイルを登録する機能の実装により，これらも満たすようにした．評価実験により，追加した機構が文献調査をする際に役立つことを確認した．

Stand-alone Full-text Retrieval System for Supporting Search of Written Materials

Yukiharu Watagami[1] Takehiko Murakawa[2] Keigo Utsunomiya[3] Masaru Nakagawa[2]
[1]Graduate School of Systems Engineering, Wakayama University
[2]Faculty of Systems Engineering, Wakayama University
[3]Faculty of Literature, Osaka Otani University

We have been developing the full-text retrieval system which is available in Internet-unconnected environments. In this paper, we report the enhancement of the system for the search of written materials. Portability, accuracy and timeliness are defined as requirements for use and we showed that our system satisfies all these properties. Moreover we attempted to improve the search interfaces, implement a unique bookmark feature, and supply the registration support in order that the system can satisfy further properties named reutilization and installation-facility. Evaluation experiments showed that the added features are useful for the information retrieval.

1. はじめに

現在，計算機での検索・閲覧できるようテキスト化された史資料を研究に活用するのは，当然のものとなっている．例えば，新たな史資料を発見したとき，インターネット上の文書を始めとする既存の情報を照合し，作者や作成年代などの推定を通じて，当該史資料の価値を確認することに活用されている．しかし，Web の検索エンジンや，ディジタルアーカイブ関連の検索サービスは，インターネットに接続できることが前提であり，必ずしもそうでない寺院や博物館などの中では利用できないという問題点があった．

そこで，筆者らはインターネット環境がなくてもスタンドアロン（単体の計算機）で動作する全文検索システムを構築することでこの問題に取り組んできた[1]．本稿では，文献調査のための全文検索が持つべき性質として，携帯性・正確性・瞬時性の 3 つを示し，筆者らが構築してきたシステムがそれらを全て満たすことを確認した．さらに，研究活動の支援となるよう，再利用性および導入容易性に配慮してシステムを改良したので報告する．

2. 文献調査支援のための全文検索

2.1 文献調査の流れ

システム構築・利用の背景として，文献調査の流れについて述べる．調査は「現地調査」と「検討」の 2 つに分けられる．現地調査では現地資料の読み解きと最低限の文献調査を行う．そこで得られた調査内容とデジタル化した現地資料を研究室に持ち帰り，検討が行われる．この際に，前回までの調査内容やその他史資料・論文から史料批判を含め，徹底的に調査を行い，結果をまとめる．

現地調査は複数回に分けて行われることもあり，その場合には現地調査と検討が交互に行われる．また現地調査でデジタル化した内容は，史料保有者の意向や研究統括者の実施方針に基づき，適正に管理（情報流出をしないように）される．

2.2 全文検索サービスの分類

全文検索サービスは 3 種類に大別できる．まず，インターネットを利用したコンテンツ検索があげられる．Google や Yahoo!などの検索エ

(c) Information Processing Society of Japan

ンジンがこれにあたる．検索対象はインターネット上のすべてのコンテンツであり，不特定多数が利用する．2 番目に，図書館の蔵書検索サービスのような，特定の用途に関して利用される全文検索サービスがあげられる．特定の用途であることから利用者が限定され，対象もサービスの提供者に依存するのが特徴である．最後にローカル検索があげられる．これは，計算機内のファイルを検索する際に利用するサービスである．本研究もこのローカル検索に位置づけられる．

ローカル検索のためのソフトウェアについて詳しく見ていく．ここ数年で，デスクトップ検索ソフトウェアが普及した．Google デスクトップ[2]や Windows Search，Beagle[3]などが知られている．Google デスクトップは今年 9 月に配布が停止された．Windows Search（Windows サーチ[4]，Windows デスクトップ サーチとも）は Microsoft が提供する，Windows XP/Vista/7 で利用可能なローカル検索サービスである．Beagle は Linux デスクトップパソコン向けに，同様の目的のソフトウェアとして開発された．これらは，結果を瞬時に得られるのが特徴である．ただし，漢文や古語，固有名詞が混じった古典籍データに関する全文検索としては十分ではなく，検索漏れが多く見られるという課題がある．

まったく別の手法で検索をするソフトウェアとして，grep コマンドを利用したものがあげられる．grep とはファイルから検索語に一致する行を出力する Unix のコマンドである．Windows 向けには，WideGrep[5]が知られている．これらは，ファイルまたはファイル群に対する逐次検索を行う．インデックスを作らない点や，正規表現で検索できるといったメリットがある一方で，検索のたびに全文を走査することになるため，テキスト量に比例した検索時間がかかること，また行単位の文字列パターンマッチングを行うので，ファイル単位の検索や AND 検索・OR 検索は容易でないといった欠点がある．ただし，全文検索の正確性検証に grep を併用するのは有用と思われ，[6][7]で使用している．

ここで，データベース管理システム(DBMS)を用いた検索システムについて述べておく．全文検索機能を持つ DBMS もいくつかあるが，日本語や漢字を含む検索については必ずしも十分な機能を有しているとは言えない．対象データや検索ニーズに特化して，データベースを設計し，システムを構築して検索に供する（例えば[8]）というのも一つのアプローチであるが，その場合，データベースに格納すべきキーワード等について十分な配慮が求められる．さらに，検索方法が非常に制限され，手軽に検索できるものとはならない．したがって，DBMS は全文検索そのものというよりは，年代など，付随データの管理や検索に利用するのが適切であると考えている．

2.3 聖教書誌情報全文検索システム

2.1 節で示した「検討」の作業を支援するものとして筆者らはこれまで，全文検索エンジンである Hyper Estraier[9]を利用した聖教全文書誌情報検索システムを構築してきた[10]．これを利用することで聖教を研究している専門家がインターネットを介して，聖教情報の検索を瞬時に行うことができる．

しかし，このシステムは Linux サーバ上で実現したためインターネットに接続できない環境では利用することができないという問題があった．聖教の原本は寺院や博物館などに保管されていることが多く，原則的に持ち出すことができない．また，原本が所蔵されている現地にネットワーク環境が整っているという保証がなく，システムを利用できない可能性があった．

この問題を解決するため，原本が保管されているその場所で検索・調査できるスタンドアロン型聖教書誌情報全文検索システム（「前システム」という）を構築した[1]．これにより，ネットワーク環境に左右されずにシステムを利用できるようになった．すなわちこれは，2.1 節で示した「現地調査」を支援するツールとなり得る．ただし，このシステムで扱うことのできる史資料を既存の聖教書誌情報に特化したため，他の関連する史資料に関して検索することができなかった．

2.4 システム要件

ここで，文献調査支援に求められるべき性質である，「携帯性」，「正確性」，「瞬時性」について述べる．携帯性とは他の計算機（サーバ）と通信せずに検索が利用できることを指す．正確性とは検索漏れがないことを指す．このとき，指定した語句（検索語）を含む文書を全て表示するようにし，不要な内容はユーザの目で排除すればよいため，検索ノイズは許容する．瞬時性とは検索時間が登録した文書数や文字数に依存することなく，1～2 秒で結果が表示されることを意味する．

この 3 つの性質を満たすかどうか，2.2 節および 2.3 節で述べたソフトウェアやシステムにあてはめると，次のようになる．まずデスクトップ検索ソフトウェアは，携帯性と瞬時性の 2

(c) Information Processing Society of Japan

点を満たすが，正確性まで満たすとは言い切れない．次に grep などの逐次検索コマンドは，携帯性と正確性を満たすが，検索が文書群のバイトサイズに比例するため，瞬時性を満たさない．そして[10]のシステムは，正確性と瞬時性を満たすが，サーバとの通信が必要となるため，携帯性は満たさない．最後に，[1]のシステムは，携帯性・正確性・瞬時性の 3 点をすべて満たす．瞬時性を支えるのは全文検索エンジン Hyper Estraier の採用であり，正確性の根拠は，N-gram に基づくインデックス構築である．また携帯性については，スタンドアロンで運用できるようにすることで，達成している．

文献調査支援を考慮に入れ，さらに実用として必要となる性質を定義する．「再利用性」と「導入容易性」である．再利用性とは，ある時点での検索結果の中から，利用者が有用と感じたデータを記録しておき，別の検索結果と見比べたり，論文に記載したりするなどして，検索結果をあとで利用できることをいう．導入容易性について，[1]や[10]では文書の登録を開発側で行っていたが，他の記録への適用やデータ保全を考えた場合，利用者自身が登録や，その前段としてソフトウェアの導入をする必要があり，そのような作業が円滑に行えることを念頭に置いている．その場合の利用者は，全文検索やデータベースの内部構造は知らないが，標準的に手持ちの PC を利用して，ブラウザを用いた検索や，ソフトウェアのインストールができる者を想定している．

したがって本稿で述べるシステムは，携帯性・正確性・瞬時性の 3 点を維持し，さらに再利用性・導入容易性の 2 点を満たすよう，[1]に機能拡張を加えたものとなる．

3. スタンドアロン型全文検索システム

3.1 システム構成

本研究ではスタンドアロンのシステムを構築するにあたり，1 台のノート PC 上で実現することとした．これは，持ち運びが可能であり，かつ漏れのない検索として全文検索を取り入れられるからである．ノート PC の他に，タブレット端末や Android 端末を検討したが，携帯性および導入容易性を満たすよう，システムを構築・利用する機器としては向いていないと判断し，対象としていない．

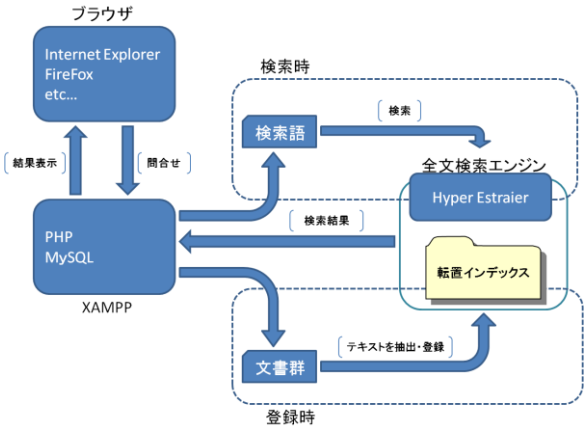

図 1: システムの流れ

登録および検索の流れを図 1 に示す．システムは， Web アプリケーションとなっている．システム構築に使用したソフトウェア[11][12]を表 1 にまとめた．システムで扱うことのできるファイルはプレーンテキストに，Word ファイル，PDF ファイルの 3 種類となっている．なお，Word ファイルや PDF ファイルからテキストを抜き出す際に，xdoc2txt を利用している．DBMS は全文検索ではなく，キーワードなどの付随的なデータを管理するために利用している．

表 1: 使用ソフトウェア

環境	ソフトウェア
運用環境	XAMPP 1.7.4
Webサーバー	Apache 2.2.17(Win32)
開発言語	PHP 5.3.4
DBMS	MySQL 5.0.7
全文検索エンジン	Hyper Estraier 1.4.10(Win)
テンプレートエンジン	Smarty 2.6.9
テキスト変換	xdoc2txt 1.37

システムが提供する主な機能として，検索機能，文書登録機能，ブックマーク機能，キーワード登録機能の 4 つが挙げられる．以下では，登録機能，検索機能，ブックマーク機能の 3 点を説明する．

3.2 登録機能

登録機能について述べる．2.4 節で触れたとおり，本システムで文書を検索するにはあらかじめシステムに文書を登録する必要があり，その作業を研究者自身が行わないといけない．そこで，容易に文書を登録できることを目的としている．登録は，ディレクトリ単位で行う．登録したい文書があるディレクトリを指定し登録することで，そのディレクトリ内にある対象の 3 種類のファイルを探し出し，インデックスを作成する．このとき，1 度登録したファイルは再度登録されることはないが，登録した後で編集されたファイルを登録した場合は，インデッ

(c) Information Processing Society of Japan

クスが更新されるようになっている．なお，インデックス作成中に他の作業をした場合，インデックスが壊れてしまう恐れがあることから，作成中はブラウザ上の操作を行えないようにしている．登録が完了すると登録にかかった時間が表示され，作業が行えるようになる．

3.3 検索機能

登録を終えると，検索が可能となる．検索フォームの画面を図 2 に示す．

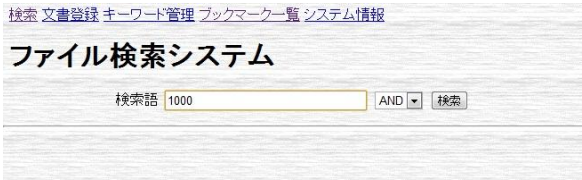

図 2 検索画面

通常の検索に加え，複数の単語で検索するには，単語をスペースで区切ることで可能となっている．このとき，AND と OR の 2 種類の検索が行えるようになっている．検索語に 3,4 桁の整数値が含まれている場合，[10]と同様に年号変換を行うようになっている．「1000」を検索語として与えた時の，検索結果画面例を図 3 に示す．結果には，文書リンク，スニペット，ファイルパスが文書毎に表示される．文書リンクとは，ファイルを直接開くことができるリンクのことで，クリックすることで該当ファイルを開くことができる．スニペットとは文書内で検索語が出現する周囲の文章のことで，文書の中でどのような場所で出てくるのかを確認できるようになっている．また，このスニペット上において，検索語の部分に黄色の背景色をつけることで表示の強調を行なっており，システムにキーワードとして登録している単語には検索リンクが生成されている．検索リンクをクリックするとその単語で検索した場合の結果を表示することができる．

検索結果の表示には，他に検索ヒット数と検索にかかった時間等も表示されている．結果は 10 件が表示されるが，それを超えた結果がヒットした場合，「次へ」というリンクが表示され，続く 10 件が表示に追加される．このとき，ページ遷移はしないため，全ての結果を同一ページにて確認することも可能になっている．

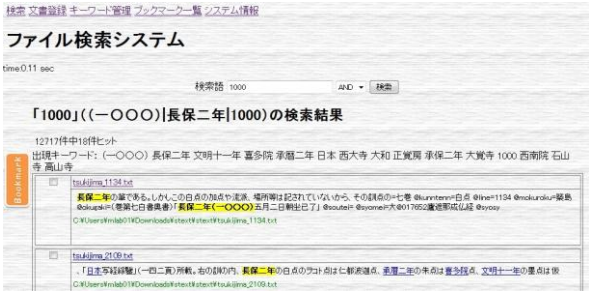

図 3 検索結果の画面例

3.4 ブックマーク機能

この機能は，検索結果の再利用を実現するための機能であり，検索結果表示から利用する．図 4 に示すとおり，検索結果から必要な文書を，左側の欄から選択する．選択した欄は色が変わり，選択状態が一目でわかるようになっている．選択が終えれば，ブックマークボタンから登録を行う．

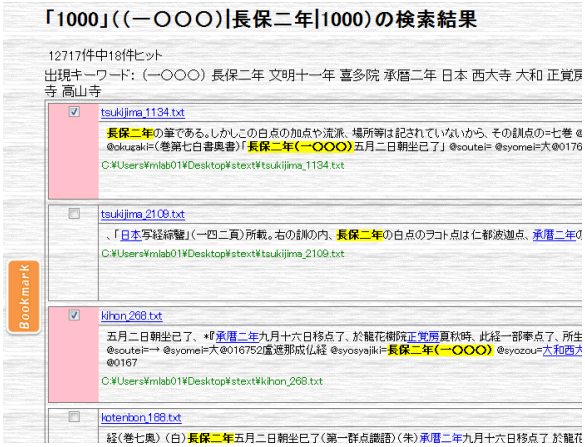

図 4 ブックマーク選択画面

ブックマーク登録画面を図 5 に示す．この画面では，登録名と考察などを入力するメモ欄が存在し，2 点を入力が終われば登録を行うことができる．

図 5 ブックマーク登録画面

登録したブックマークはシステムメニューから一覧として確認することができ，再編集が可能となっている．考察が終わったブックマークについては削除することもできる．

(c) Information Processing Society of Japan

4. 評価実験

4.1 実験方法

本システムを評価するにあたり，2 種類の実験を行っている．1 つは「導入実験」であり，研究者がシステムを実際に利用できる状態にできるかを検証する．これは，本システムがこれまでの開発者ベースのシステムではなく，研究者ベースのシステムとなっており，利用者（研究者）自身の手によってインストールができる必要があるからである．もう 1 つは「利用実験」であり，本システムを実際に使用して様々な機能に対して使いやすさを検証する．

4.2 実験結果

まず「導入実験」の結果について述べる．実験は，和歌山大学内の研究室および人文研究者（宇都宮ほか 2 名）で実施し，研究室内ではデータとして [10]で述べた聖教書誌情報のうち，出典目録が東寺のものを除く約 12,000 件を使用した．システム導入時に，TeamViewer とポート番号の競合による問題が発生したが，TeamViewer が 80 番ポートを使用しないようにすることで回避した．Hyper Estraier を利用して全文検索を行うにはあらかじめ転置インデックスを作成しておく必要がある．この登録にはネットブック（OS:Windows XP 32bit, CPU:Atom 1.6GHz, メモリ:1GB）で上述の約 12,000 件に対して，約 25 分かかり，インデックスサイズは 19.5MB となった．

これに関係して，登録数と容量の増加によるインデックス登録時間の変化について実験した結果についても述べる．実験には，Wikipedia の文書ファイルを利用した．登録数の増加による登録時間の変化を図 6 に示す．この実験では，登録する容量を 25MB に固定し，ファイル数を 1～1000 で変化させている．ファイル数が増えれば登録時間が伸びていることが見てわかる．次に，容量の増加による登録時間の変化を図 7 に示す．この例では，ファイル数は 100 で固定し，容量を 1MB～100MB まで変化させた．このように，登録時間は登録数と容量に比例することがわかる．

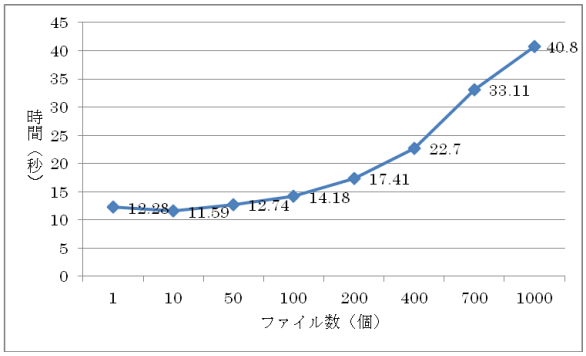

図 6 ファイル登録時間（容量固定）

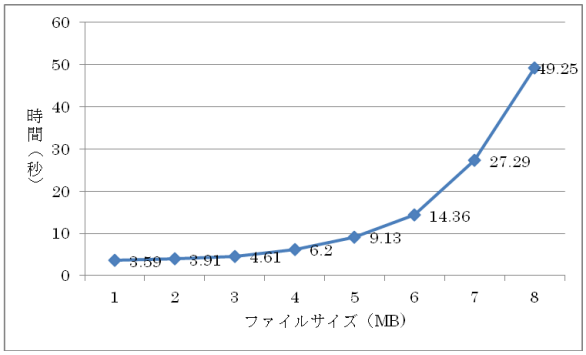

図 7 ファイル登録時間（登録数固定）

続いて，「利用実験」の結果について述べる．「導入実験」により導入した本システムを 1 ヶ月間利用してもらった後，検索と登録の機能について評価をいただいた．この時，調査現場における携帯性を満たすシステムの有用性と，文献から得られた様々な考察・意見を登録することができるブックマーク機能について高評価を得ることができた．その上で，システムをより有用なものとするための改善点をいくつか挙げられた．

4.3 考察

「導入実験」の際に得られた意見から，登録機能の問題点があげられた．登録する際，本システムでは 1 つのファイルに 1 つの文書といった対応付けが必要となっている．しかし，研究者の扱うファイルの中には，Excel ファイルの 1 行に 1 つの文書データを扱うことで複数の文書データが 1 つのファイルに収められているものが存在する．システムでこういったファイルを利用するには，分割して保存しなければいけなくなり非常に手間となってしまう．このことから，システムの利便性が損なわれてしまうというのが問題点としてあげられた．対応案として，登録の際に，内部的に文書を分割して登録する機構を実装することが考えられる．

次に，「利用実験」から得られた意見から，

(c) Information Processing Society of Japan

システムの改善点について述べる．文献調査の都合上，検索の際，段階的な検索の絞り込みや検索履歴の利用が行えることが利便性において必要であると考えられる．また，ブックマーク機能に関し，考察の内容が複数にわたることがあることから，タグ付けや項目毎のメモの登録機能などが考えられる．なお，ブックマーク機能が評価された理由としては，従来のブラウザなどによるブックマークは，URL や検索ページ自体の登録であり，必要な文献のみを絞り込むような利用は出来なかったため，「再利用性」の面で実用的ではなかったことがあげられる．

なお，本システムはノートパソコンだけではなくデスクトップパソコンにも導入することができる．よって，ノートパソコン等では扱い切れないような大規模の文書のインデックス化や検索も技術的には可能である．

5.おわりに

開発中の全文検索システムについて，その機能と特徴，既存関連システムとの違い，評価実験の現状について述べた．

本システムを利用することで，全文検索の内部構造など詳細な知識を必ずしも有しない人文研究者が，Web の検索エンジンを利用するのと同じ使いやすさで，手持ちの文書群から全文検索ができるようになる．国内の中世の文献に限定したとしても，史資料に対するテキストは多種多様であり，漢文や古語のほか，口語でメモが書かれることもある．Hyper Estraier の N-gram に基づくインデックス作成により，形態素解析を不要とし，複数の言語であることを意識することなく検索が可能となっている．

今後の課題として，さらなる機能改良と実際の現時調査での利用評価，そして USB フラッシュメモリにシステムおよびデータを収納したポータブル版の開発・提供などが挙げられる．

謝辞　評価にあたり，龍谷大学 野呂靖先生ならびに大谷由香先生に使用していただくとともに，ご助言を賜りました．ここに感謝します．

参考文献

[1] 渡上将治, 村川猛彦, 宇都宮啓吾, 中川優:聖教書誌情報全文検索システムのスタンドアロン化の試み，情報知識学会誌，Vol20，No.2, pp.189-194, 2010
[2] Google デスクトップ: http://desktop.google.com/
[3] Beagle: http://www.beagle-project.org/
[4] Windows Search – Windows 7 の機能：http://windows.microsoft.com/ja-JP/windows7/products/features/windows-search
[5] WideGrep: http://homepage2.nifty.com/AnoSoft/
[6] 村川猛彦, 丁敏, 中川優: 仏典全文検索システムの構築と評価，人文科学とコンピュータシンポジウム論文集，情報処理学会シンポジウムシリーズ，Vol.2007，No.15，pp.221-228, 2007.
[7] 河中健馬, 渡上将治, 村川猛彦, 中川優: 全文検索システム選定支援システムの構築，情報知識学会誌, Vol.21, No.2, pp.191-196, 2011.
[8] 木村文則，小牟礼雅之，前田 亮，佐古愛己，杉橋隆夫：古典史料データベース検索システムの提案，情報処理学会研究報告 人文科学とコンピュータ，2008-CH-78，pp.45 - 52 (2008).
[9] 全文検索システム Hyper Estraier: http://fallabs.com/hyperestraier/
[10] 村川猛彦，宇都宮啓吾，中川優: 聖教書誌情報全文検索システム，情報処理学会論文誌: データベース, Vol.1, No.3, pp.96-107, 2008.
[11] XAMPP: http://www.apachefriends.org/jp/xampp-windows.html
[12] xdoc2txt http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html

(c) Information Processing Society of Japan