社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－170 (6)
2005／11／21

利用者のプロフィールを考慮した連想検索 OPAC の構築

當山仁健 長田智和 玉城史朗

toyama@dit.ic.u-ryukyu.ac.jp {nagayan, shiro}@ie.u-ryukyu.ac.jp
琉球大学理工学研究科情報工学専攻 琉球大学情報工学科

図書館の OPAC で従来用いられているブール代数を用いたキーワード検索では検索の精度はキーワードの選択に大きく依存するため，調べたいものに対する利用者の知識が不十分な場合，検索洩れが発生しやすい．また，調べたいものに対するイメージが漠然としている場合などには欲しい情報を取り出すことは難しい．この問題を解決し，高精度な利用者指向の OPAC を実現するため，類似性による連想検索の技術を用いた OPAC を実装した．また，検索の入力は同じでも，利用者のプロフィールによって要求する情報は異なる場合が多い．利用者の背景を考慮して求めている検索結果を提案するため，その利用者の属するグループ (学科) の図書の貸出状況を評価尺度に加味するシステムを作成し，試験運用した．

Associative Searching OPAC considering User's Profile

Yoshitake TOYAMA Tomokazu NAGATA Shiro TAMAKI

toyama@dit.ic.u-ryukyu.ac.jp {nagayan, shiro}@ie.u-ryukyu.ac.jp
Dept. of Information Engineering The University of the Ryukyus

The precision of the search result with OPAC used in most of the library, which uses keyword search system with boolean algebra, depends largely on the keywords of the input. To retrieve better result, you need to use appropriate keywords. However if a user has little knowledge about what he want to retrieve, it is difficult to choose appropriate keywords. The precision of search result depends on the knowledge of what they try to search. To solve this, we built associative search OPAC.
Even if a user used same keywords as others, the need of each user varies. To offer result satisfying each user's need, we implemented a system which consider user's profile(the department of the user) to make effect on the search result retrieved with associative search.

1 はじめに

IT の発展により，大量のテキストから目的の情報を検索することが可能になり，図書館では書誌事項からコンピュータを利用した検索が可能になった．ほとんどの大学図書館等では自館の目録をオンライン化し，OPAC(Online Public Access Catalogc) として利用者に提供している．その際，必要な情報へアクセスするための技術として利用されることが多いのは，ブール代数 (AND/OR 等) を用いたキーワード検索システムや，NDC などを基に資料を分類したディレクトリ型の検索である．このようなシステムでは検索の精度はキーワードの選び方や前提知識に大きく依存するため，大規模，高精度の検索システムを実現するのは難しい．

それに対して自然文を入力キーとして類似性を元に検索を行う連想検索技術を用いれば，検索漏れを防ぎ，利用者の知識が不十分なものについてもある程度検索することが可能である．このような連想検索の効果は以前から知られていたが，計算量などの問題で実用レベルでのサービスの実現は難しかった．しかし，近年連想検索を実装するためのツール [1] が公開され，連想検索を利用したサービスはさまざまな分野で実用されはじめている [2, 3, 4, 5]．

また，同じキーワードを用いて検索を行なっても，利用者の興味やバックグラウンドによって求められる情報は異なることが多い．さらに，同じ内容を取り扱っ

	文書1	文書2	文書3
単語1	3	0	6
単語2	0	3	7
単語3	2	0	0

図1 索引語-文書の行列

た資料でも内容的に評価の高いものとそうでないものがある．これらの見極めは探している情報に関する知識が十分でないと難しい．そこで本研究では，検索漏れを少なくするため，まず，連想検索を用いたOPACを構築し，評価尺度に利用者のプロフィールに基づいた重みを加味した検索システムを構築した．

2 関連技術

本システムで利用する各種技術について説明する．

2.1 連想検索

2.1.1 ベクトル空間モデル

図1のように，索引語を行とし，出現文書を列とする行列を定義し，行列の要素を各索引語の重みとする．各列を列ベクトルとみなすと，文書間の類似度はベクトルの類似度とみなすことができる．そしてベクトルの類似度は内積を計算することで求められる．これにより類似度を元にした連想検索を実現することができる．

この手法の有効性は以前から知られていたが，文書の数が増えるほど，次元が増加し，計算量は2乗オーダーで増加するため，実現は難しかった．しかし，この計算を効率よく実行できるツール [1] の公開により，大規模なデータに対してもこの手法を利用した連想検索環境が構築できるようになった．

2.1.2 索引語に対する重み付け

図1での行列の要素である重みとして，単純に出現頻度を用いると，出現頻度が高いというだけで一般的過ぎる単語も検索対象になってしまい，ノイズが多くなってしまう．そこで，実際にベクトル空間モデルを利用する際にはそのような単語の重みを下げ，特徴的な単語の重みを上げるためのアルゴリズムを通して重みを計算し，それを行列の要素として利用する．代表的なアルゴリズムに式(3)のようなtf・idf法がある．tf・idf法とはある文書(d)での単語(t)の出現頻度(Term Frequency．式(1))と全文書中(N)でその単語が出現する文書の数(Inverse Document Frequency：単語の特定性を表す．式(2))を掛け合わせたもので，全体の文書の中では出現頻度は低いが，特定の文書の中で出現頻度が高い特徴的な単語の重みを上げ，多くの文書に数多く出現する単語の重みを下げるアルゴリズムである．その他にtf・idf法よりも精度が高いSinghalの方法がある [6]．

$$w_t^d = tf(t,d) \tag{1}$$

$$w_t^d = idf(t) = log\frac{N}{df(t)} + 1 \tag{2}$$

$$w_t^d = tf(t,d) \cdot idf(t) \tag{3}$$

2.2 形態素解析

図1のような行列を作成する際には，まず文章から索引語となる単語を選び出す必要がある．英語などのように，単語間の区切りがはっきりしている(空白で分割されている)言語に関しては，単純に処理できるが，日本語のように各単語が独立していない膠着語の場合，品詞ごとに切り出し，形態素に分割する必要がある．

2.3 適合性フィードバック

検索の際，一度検索条件を入力をするだけで，目的の情報にたどり着くことは現実的には難しい．システムの提案する検索結果に対して利用者が評価を行い，評価結果をシステムにフィードバックし，検索を繰り返させることで精度を上げることができる(適合性フィードバック)．連想検索を利用すると，システムが入力から単語を連想して検索者に提示することができる．その中から利用者が自分の興味の方向に近いものを選ぶことで，検索結果の精度を挙げることができる．また，検索の途中で興味の方向が変わった際にも柔軟に軌道修正をすることが可能である．

2.4 利用者モデル

表1のように，検索語は同じでも検索者が求めている内容は異なることは多い．また，同じものごとについて書かれた資料でも，評価の高いものとそうでないものもある．初学者にはこの見極めが難しい．そこで検索者をモデル化し，同種の情報を求めているグルー

プの評価を利用することで，高精度の検索結果が期待できる．

表1　所属学科による求める情報の違い (例)

所属学科	検索語	求めている情報 (例)
人間福祉学科	アルコール飲料	アルコールの依存性について
産業情報学科	アルコール飲料	産業としてのアルコール飲料について
社会文化学科	アルコール飲料	文化としてのアルコール飲料について

3　利用者のプロフィールを考慮した連想検索 OPAC

高精度な利用者指向の OPAC の実現のため，前述の連想検索，適合性フィードバック，利用者のプロフィールによる重みを取り入れた OPAC を構築し，試験運用した．

3.1　システム構成

システムの流れは図2のようになっている．また，システムの構成は図3のようになっている．以下に処理の流れを示す．

1. 所属する学科を選択する．その際，ログインした日時と学科をログとして記録する．
2. 検索語を入力する．入力を元に連想検索を行い，入力と関連する書誌を 200 件取得する．
3. 連想検索の結果としてシステムにより選ばれた書誌に対して，利用者の該当する学科の貸出状況を重みとして取得．
4. 取得した重みを連想によって算出されたスコア (類似度) に掛け合わせる．利用者に結果の書誌のリスト (30 個) と，連想によって選ばれたキーワードを提示する．
5. 利用者は提示された書誌リストの中に満足いくものが無ければ，同時に提示されるキーワードのリストから関連するキーワードを選び，満足のいく結果が得られるまで，システムに入力として渡して再検索する (適合性フィードバック)．
6. 利用者は興味のある書誌を選択する．この際，適合性フィードバックの回数と入力として渡された検索語を記録する．
7. 内容に関する情報などから，求めている情報が記載されているか，利用者はアンケートに答える．

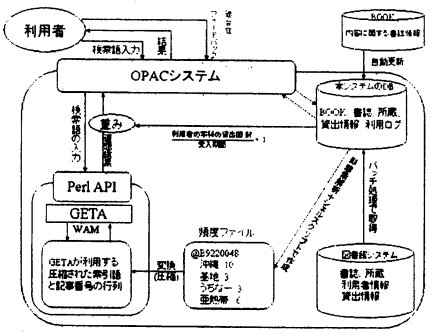

図3　システムの構成

3.2　連想検索 OPAC の実装

連想検索をするための類似度の計算は汎用連想計算エンジン GETA [1] を用いて実装した．GETA に与える索引語と文書名の行列の外部表現である頻度ファイルを生成するために，日本語形態素解析システム ChaSen [9] で形態素解析を行った．その際，ストップワードとみなせるような品詞はフィルタリングを行なっている．また，沖縄の独特な言葉や固有名詞に対応するため，沖縄辞書 [10] を用いて辞書を強化した．

3.3　利用者のプロフィールによる重みづけ

当初，本システムでは利用者のプロフィールを考慮した重みとして，所属する学科と年次でグループ分けを行い，それぞれのグループの貸出状況を重みにすべく，システムを実装した．しかし，全資料数に対して学生の貸出数が少ないため，連想による検索結果に重みとしての影響を与えていることが体感できなかった．そこで利用者グループの粒度を大きくするため，学年は気にせずに，学生の所属している学科を重みに利用した．

現在，特定の学科での書誌ごとの貸出件数 (r) をその書誌の一番古い所蔵の保持年数 (α) で割り，1 を足したもの (w) を体験的に利用者のプロフィールで算出した重みとしている．これを連想によって引き出された重みに掛け合わせることによって，検索結果のスコアへ利用者のプロフィール情報を反映される．

$$\alpha = \frac{today - k}{365} \tag{4}$$

$$w = \frac{r}{\alpha} + 1 \tag{5}$$

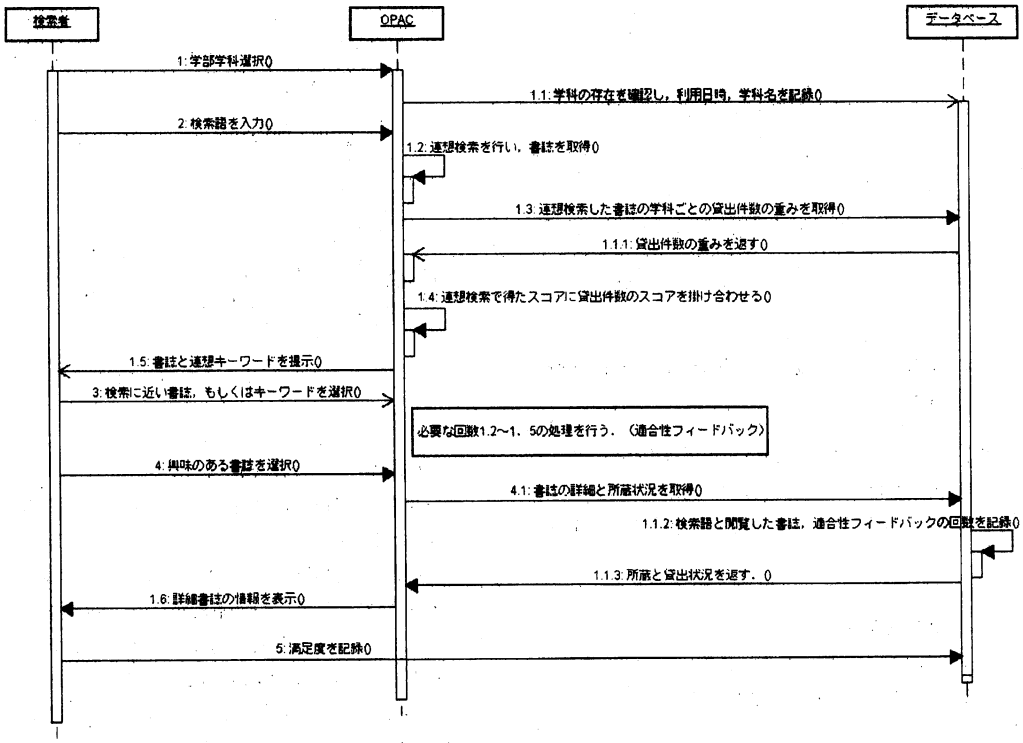

図 2　システムの処理の流れ

3.4　システム環境

実際にシステムで利用している環境は以下のとおりである．

ハードウェア構成

- CPU: Pentium4 2.2GHz
- Memory: 640MB
- HDD: 40GB

ソフトウェア構成

- OS: Fedora Core 4
- 汎用連想計算エンジン GETA
- 日本語形態素解析システム ChaSen
- Apache 2.0.54
- PostgreSQL 8.0.4
- Oracle 8.1.7(図書館システム)
- 内容に関する書誌事項 日外アソシエーツ BOOK [12]

3.5　データベース構成

プロフィールの重みと取得している履歴データに関するデータベース構成は図 4 のようになっている．図書館システムで利用している既存の書誌事項 (NACSIS-CAT [11]) と内容に関する書誌事項 (日外アソシエーツ BOOK) は ISBN で結合した．また，ログに関しては以下の項目を保存している．

- 利用者の利用開始時間
- 利用者の所属する学科
- 検索語
- 適合性フィードバック回数
- 入力した検索語ごとの参照した書誌
- 利用者の検索に対する満足度

3.6　インタフェース (適合性フィードバック)

インタフェースは適合性フィードバックが実現できるように作った (図 5)．画面左側には連想によって引き出された書誌リストにプロフィールの重みをかけた

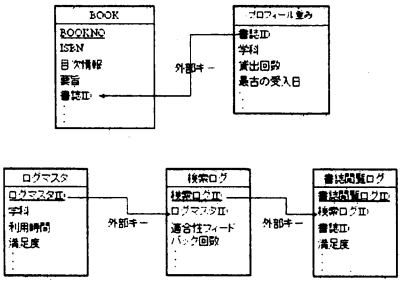

図4　データベース構成

スコア順に30個表示する．画面右側には連想によって引き出されたキーワードのリストが30個表示される．利用者は右側のキーワードから自分の興味に近いものを選び，システムに再度フィードバックすることでさらに精度の高い検索結果を得ることが可能である．自分の興味のある資料を無制限に選択し，システムに入力することで内容が似た資料のリストを得ることも出来る．また，ステータスバーには書誌に関する内容を表示することで画面遷移数を減らしている．

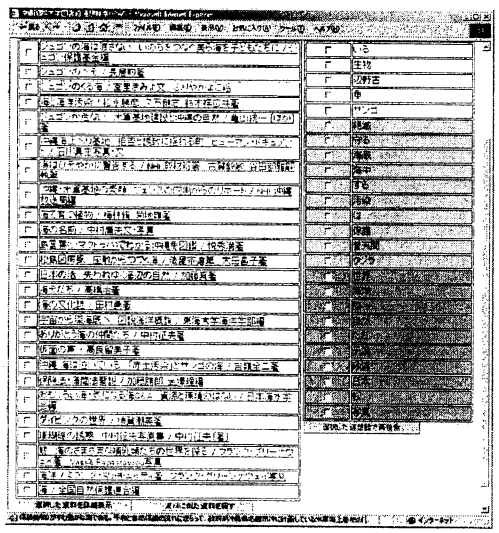

図5　書誌リストとキーワードリストの提示

4　評価

2005年10月18日から2005年10月28日まで図書館内に4台のクライアントを設置し，ログを取得した．実際にクライアントで検索語を入力して検索された件数は970件であった．その内アンケートに回答したのは43人．総利用者の4.4%であった．アンケートの結果は図6のようになった．それぞれの利用者が行なった適合性フィードバックの回数は図7のようになっている．

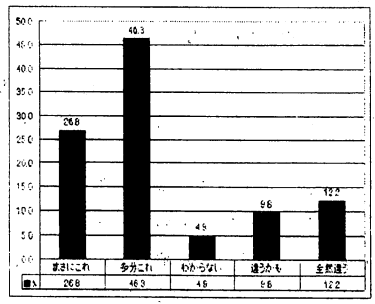

図6　満足度アンケート

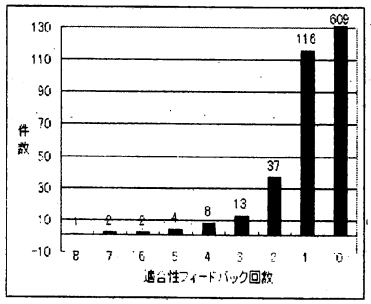

図7　適合性フィードバックの回数

4.1　考察

図6によると，「まさにこれ」「多分これ」と回答したのは全体の73.1%であり，「違うかも」「全然違う」を選択した利用者は全体の22%にすぎず，多くの利用者が満足を感じて検索を終了していることが読み取れる．

図7をみると，実際に適合性フィードバックを利用して繰り返し検索している利用者が少ないことがわかる．これは従来のキーワード型検索に慣れている利用者にとって，適合性フィードバックを利用するという

発想が無いためだと思われる．また，実際にログを確認したところ，文章を入力して検索している利用者は数名しかいなかった．キーワードを入力している利用者がほとんどであった．やはり，キーワード型検索に慣れた利用者にとっては，自然文で検索できるというのは未経験なため，抵抗があるようだ．今後，利用者教育を充実させて利用方法を周知することで，さらに有効にシステムを活用できるようになるのではないかと考える．

5 まとめと今後の課題

本システムで取得している利用者が入力した検索語と，実際に閲覧した資料の関係性をシステムの評価尺度に組み込むことにより，さらに高精度のシステムを作りたい．また，今回は学科毎で利用者をグループわけしたが，さらに詳細な利用者情報，たとえば読書遍歴などを考慮してシステムの評価尺度に応用できるシステムを作成したいと考えている．今後は利用者のプロフィールによる重みの効果の検証方法についても検討していく．

参考文献

[1] 高野明彦，西岡貞吾，丹羽芳樹．連想に基づく情報アクセス技術．情報の科学と技術，54 巻 12 号，2004.

[2] Webcat Plus. http://webcatplus.nii.ac.jp/

[3] 文化遺産オンライン．http://bunka.nii.ac.jp/jp/

[4] 新書マップ．http://shinshomap.info/

[5] ASCII 24. http://ascii24.com/news/

[6] 高野明彦，西岡貞吾，今一修，岩山真，丹羽芳樹，久光徹，藤尾正和，徳永健伸，奥村学，望月源，野本忠司．IPA2001 年度成果報告論集，http://geta.ex.nii.ac.jp/pdf/itx2002.pdf

[7] 汎用連想計算エンジン (GETA). http://geta.ex.nii.ac.jp/

[8] 徳永健伸．情報検索と言語処理，東京大学出版会．

[9] 日本語形態素解析システム ChaSen 「茶筌」. http://chasen.aist-nara.ac.jp/

[10] 沖縄辞書．https://sourceforge.jp/projects/o-dic/

[11] NACSIS-CAT. http://www.nii.ac.jp/CAT-ILL/contents/home.html/

[12] 日外アソシエーツ BOOK. http://www.nichigai.co.jp/database/book-plus.html