グループウェア 35−17
（2000. 3. 10）

関連文書によってフィルタリングする連想方式情報検索ツールの開発

李 健　金井 貴　國藤 進
北陸先端科学技術大学院大学　知識科学研究科

要旨： WWW の普及が一般化にしつつある現在，WWW における情報検索には，検索の有効性と効率性がその検索システムの性能を評価する基準となっている．これを目標にして，今までさまざまな検索システムが開発されてきた．本研究は，本学の国藤研究室で開発された連想方式情報検索システムにおける適合率の改善を目指すため，ユーザが作成した文章やウェブの情報などに基づいて抽出した関連文書をプロファイルとして利用する上で，名詞キーワードの頻度における文書の意味情報を着目し，高頻度語に語の出現頻度に関するモデルやベクトル空間型モデルなどの手法を使って，システムにフィルタリング部を実装した．このフィルタリングと関連語を用いた検索により，一つの検索システムに再現率と適合率の両方を改善するプロセスが共存し，且つ協調的に動作することが出来る．

An Information Filtering based on Related Document for an Association Model Information Retrieval System

Jian LI　Takashi KANAI　Susumu KUNIFUJI
School of Knowledge Science,
Japan Advanced Institute of Science and Technology, HOKURIKU

Abstract: In this paper, we propose a new filtering model for "Contents Retrieval System of Academic Researchers" which has been developed in JAIST. Through comparing similarity metrics between the related document and objective files, the filtering system improved the precision of the system. Since the information retrieval system was built to respect for the recall of information retrieval, we also propose a new approach that is to make two processes which service for the recall and the precision respectively to work collaboratively in the same information retrieval system.

1. はじめに

Web の隆盛に伴って，現在，公的，私的な情報を家にいながらにして簡単に手に入れることができる社会になった．これまでの数年間のうち，Web 上での情報検索に関する研究や情報検索における商用サービス等は，盛んに行われるようになっている

しかしながら，情報洪水となりつつある現在，検索エンジンの有用性が徐々になくなりつつあり，玉石混淆の検索結果に対して，目的のサイトを絞り込むのに膨大な時間がかかってしまう傾向が益々強くなっている．混沌とした情報空間の中で，素早く手軽に検索したい情報を手にするには，検索エンジンを改良するだけではなく，情報フィルタリング技術も必要となる．

従来，ネットワーク上で生じるテキスト情報のフィルタリングに関する研究は，主に Web が普及する前から USENET 等のグループコミュニケーションシステムでよく行われてきた．WWW や USENET 等を対象とした情報検索の研究は，1994 年前後に盛んに研究され，SIFT や GroupLens 等の二十数種のフィルタリングシステムが開発されている[1]．

WWW における情報検索には，検索の有効性と効率性がその検索システムの性能を評価する基準となっている．これを目標にして，今までさまざまな検索システムが開発されてきた．

本研究では，本学の国藤研究室で開発されている連想方式情報検索システム（AMIRS：Association Model Information Retrieval System）の後継開発の一環として，フィルタリング部をシステムに実装することで，このシステムにおける適合率問題などに対処する[2]．

2. 連想方式情報検索システム

北陸先端科学技術大学院大学において，石川県下の研究者に関するデータベースを構築し，研究者を関連するキーワードから検索する方法として開発されている大学研究者の意味探索システムは，連想方式情報検索システム(AMIRS)と呼ばれ，データベースに入っている名詞単語間の距

離尺度を顧慮したキーワードベクトルを用いて関連語辞書と構築したことにより，キーワード検索と同時に関連語の提示による連想検索も可能となっている．従って，このシステムでは，企業等で使用されている一般的な用語と大学等の研究機関における専門用語の解離に対して，システムによるキーワードの連想により，企業ニーズと研究シーズをお互いに関連付けることが出来る．連想辞書をベースにした連想方式情報検索を利用することで，検索者の持つ知識の程度にかかわらず有用な情報を検索できるという面では，このシステムは従来のキーワード検索方法より検索の再現率を改善している．

しかし，現在の連想方式検索システム AMIRS における頻度情報辞書は，全体のキーワード間のスコアは正規化されていないため，キーワードの増減によって探索結果の順位が大きく変わってしまう欠点がある．また，複数の分野の文章群にまたがってキーワード間の関連度を抽出する場合，分野によって異なる意味や使われ方があるため，連想方式検索の結果に適合率が低くなるという問題もある．

3. フィルタリングへのアプローチ

情報フィルタリングの目的は，ユーザの興味や関心を記述したプロファイルを参照して，情報源から次々と流れてくる情報のうちユーザの関心があるものだけを取り出すことである．ユーザの情報要求を満たす情報をユーザに提供するということから，フィルタリングは，情報検索の関連技術の一つであり，情報検索におけるほとんどの技術と手法を利用することが出来る[3]．

本節では，フィルタリングを実装する時に使われた，語の出現頻度に関するモデルとベクトル空間型モデルの二つの方法だけを取り上げる．

3.1 tf・idf モデル

語の出現頻度に関する最も基本的な量は，各文書内での出現頻度 f である（文書 d において語 t が出現する回数）．

各文書内での語の出現頻度は，tf・idf における tf の計算に用いられることが多い．tf・idf の考え方は，「重要な語ほど文中で繰り返して用いられる」という Hans Peter Luhn の仮定に基づく[4]．尚，idf については，$1/n$ のような単純な形式ではなく，次のような算出法が採用さるることが多い．

$$\mathrm{idf} = \log \frac{\mathrm{N}}{n} \qquad (1)$$

ここで，N は文書の総数で，n は語 t が出現した文書の数である．

ある文書の中では，tf が tf・idf の値に大きく影響を与えるが，idf をかけることによって，tf・idf の値は文書全体における語の出現頻度が高いほど小さくなるのは，この算出法の狙いである．言い換えれば，tf・idf の基本的なアイデアは，多くの文書に現われる高頻度語の重要度を下げ，少数の文書にのみ現われる低頻度語の重要度を上げるということである[5,6]．

3.2 コサイン尺度モデル

コサイン尺度（cosine measure）モデルは，三角の類比計算法と似ているから名前が付けられた一種のベクトル空間型モデルである．基本的な考え方は，ベクトルモデル空間において，全ての文書をベクトルで表現し，二つの文書の類似度を文書ベクトルのなす角の余弦で測ることである．従って，各文書を語の重み付けのベクトル値によって以下のように表現できる．

$$V = (wt_0, wt_1, \ldots, wt_n)$$

各 wt_i が語と重みを持つことにより，各文書 d_i と関連文書 q 間の類似度は，次の式で算出する[7]．

$$S(d_i, q) = \frac{\Sigma_t w_{q,t} w_{d,t}}{\sqrt{\Sigma_t w_{q,t}^2} \sqrt{\Sigma_t w_{d,t}^2}} \qquad (2)$$

式中，$w_{x,t}$ は文書 x の中で語 t の出現する次数である．

4 名詞における文書の意味情報

単語の出現頻度を用いて文書の意味を把握するには，文書の大意を表す重要な語（以下重要語とする）の抽出は不可欠となる．重要語の抽出法として，tf・idf のような計算法もしばしば使われ，よい結果が得られている[8]．また，手がかり語等を利用する方法も研究されている[9]．

しかしながら，これらの手法は，単語の頻度と文書における曖昧な表現間の関連については，考慮していない．そこで，複数の意味を持つ単語が検索に与える結果を調べるため，AMIRS の名詞データベースおよびキーワードデータベースを用い，次のような予備実験を行った．

4.1 予備実験

予備実験では，「組織」や「モデル」や「社会」等のような曖昧な言葉を使って行った．これらの語は，複数の異なる意味を持つため，キーワードとして選られた．

まず，上記の言葉をキーワードとして入力し，AMIRS から上位 10 件の結果を得る．

次に，得た検索結果に従って，データベースの中にある元のファイルデータから次のように定義する用語でファイルごとに統計分析をまとめる．

重 要 語：研究者ファイルの中で研究分野を表す名詞．

高頻度語：出現回数3より大きい，かつ上位順から15%に入っている名詞．

低頻度語：1回だけ出現した名詞．

実験を行った結果，表1のような結果が得られた（ここでは，キーワード「組織」で検索したものに対する結果のみを取り上げる）．

コード番号	キーワード全体の個数	重要語の例	重要語の出現回数		
			高	中	低
189	25	材料力学	1回		1回
105	107	材料力学 バイオメカニクス	1回	1回	
615	19	地域物産物 新品種開発・増殖	3回	2回	2回
89	120	医用生体工学 人工臓器	4回	1回	
231	41	工作機械 生産システム		1回	3回
558	82	植物育種	2回		
107	105	高分子化学 触媒化学	1回	1回	1回
120	55	熱工学	1回		1回
210	56	応用物性		1回	1回
228	58	材料力学	1回	1回	

注：コード番号は、AMIRSに登録されている研究者情報ファイルの番号を指す

表1．キーワード「組織」の検索例の結果統計

表1の結果から見ると，名詞の頻度は各文書の長さによって変わるが，文書の意味情報は高，低頻度語に多く含まれていることがわかる．

さらに，プロファイルとして使われる関連文書の名詞頻度情報におけるフィルタリング効果，すなわち，関連文書の高，低頻度語情報を使って，関連文書で伝わっている意味情報が検索結果の中の無関係となるものを削除，或いは検索結果を並び替えることが出来るかどうかの検証を行った．実験の手順は，次のようになる．

① 日経サインスの WWW サイトから取った五つの文章（コンピュータ関連，医学・生物化学関連）と本学材料研究科から取った一つの文章，合計六つの文章を関連文書に使う．

② 関連文書から切り出した名詞高頻度語と低頻度語を使って，各検索結果ファイルの高頻度語と低頻度語にそれぞれマッチさせる．

③ マッチした名詞の個数を高頻度語と低頻度語ごとに足し算し，それによってキーワードで検索されたファイルの順位を並べ替える．その場合，高頻度語を優先的に考慮され，マッチした語の数は0であれば，低頻度語によって新しい順位を決める．

④ 検索されたファイルが関わっている研究分野を参照し，関連文書によりフィルタリングする結果を考察する（表2）．例を挙げると，医学・生命科学類の関連文書2と4によって，キーワード「組織」で検索された上位10件の結果のうち，85番と105番の研究者ファイルだけが高いスコア値が与えられ，新しい並べ順ではトップになっている．これらの研究者の研究分野を調べれば，結果は適切であることが分かる．

考察の結果として，次のようなことを考えられる．

① 高，低頻度語両方を使った場合，間違い及び漏れミスは結果全体の約9%に留まっている．

② 高頻度語の結果の中に，誤りがない．

③ 高，低頻度語を両方使えば精度が高くなるが，マッピングする時間の限度があるので，高頻度語を優先的に使う．もし，こうした場合良い結果が出なければ，低頻度語を考慮する必要がある．

4.2 名詞高頻度語

予備実験の結果に基づき，連想方式情報検索システムには，名詞高頻度語が重要な意味情報を持ち，第一章で述べたフィルタリング方法に適用すれば，より効果的なフィルタリングが期待できる．

本研究では，新たに名詞の頻度情報を記載するテーブルを作成し，主に名詞高頻度語によるフィルタリングをシステムの上に実現する．

具体的には，名詞高頻度語のみに対して，tf・idf 手法，コサイン尺度モデルとマッチされた高頻度語の数によるフィルタリングする方法を三つのプログラムで実現し，それぞれのフィルタリングの効果を検証する．

					( )
			合理	誤り	漏れ	調節可

				標準回答（参考）
関連文書	1	高頻語		なし
		低頻語	~~89~~	なし
	2	高頻語	105 89	89, 105
		低頻語		89, 105
	3	高頻語		なし
		低頻語		なし
	4	高頻語	89 105	89, 105
		低頻語	89	89, 105
	5	高頻語		なし
		低頻語		なし
	6	高頻語	228 89 105 107 189 210	107, 105, 89, 210, 228, 189
		低頻語	89 (228) ~~(558)~~	107, 105, 89, 210, 228, 189

表2．組織」によって検索された例におけるフィルタリングする結果

5. 関連文書によるフィルタリング部

通常の情報検索と異なり，本研究では，クライアントからユーザのプロファイルとする関連文書をサーバへ送る必要がある．キーワード入力の枠の下にテキスト入力フォームを付加した．ユーザはコピーと貼り付けによって，検索要求に応じてWWWサイト等にある関連文書の内容を簡単に変えることが出来る（図1）．

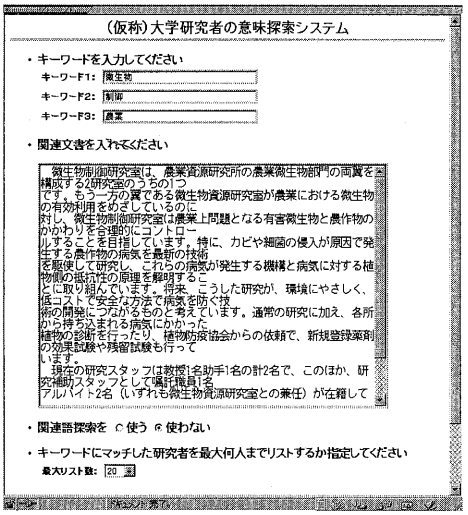

図1. フィルタリング実装後のメインページ

前節述べたように，連想方式情報検索システムにフィルタリングを実装する時，三つのプログラムを用意した．

プログラム①　関連文書から名詞を取り出して，名詞リストを作る．その中からさらに上位15%のものを取り出して高頻度語として使う．ここでは，高頻度語の定義は予備実験と少し意味が変わるが，関連文書の長さが一定の場合，両者間に大きな差がない（上位15%を取る時，関連文書の長さは400〜600字である）．このように決められた高頻度語を用いて入力キーワードで検索されたファイルと照合する．そして，照合された名詞に対して，キーワードで検索されたファイルの出力順をtf・idfの算出値でソートする．

プログラム②　関連文書の名詞リストから上位15%の高頻度語を抽出し，入力キーワードで検索されたファイル中の出現頻度が3より大きい名詞と照合する．もし，入力キーワードで検索されたファイルの名詞リストの最大頻度が3より小さければ，それを全部照合する．そして，照合された名詞の個数で出力順をソートする．

プログラム③　①，②と同じように，関連文書から取り出された名詞リスト中上位15%の高頻度語と入力キーワードで検索されたファイルの名詞リストをそれぞれ抽出し，照合させる．そして，式2に基づいて，各入力キーワードでファイルと関連文書間の内積を求める．最後に，内積値で出力順をソートする．

6. 評価実験と考察

連想方式情報検索システムにおける再現率と適合率，および三つの手法を用いてフィルタリングすることによる再現率と適合率の変化を測定するため，ラテン方格法の原則を踏まえて評価実験と考察を行う．

実験に影響する要因として，フィルタリングの効果に加えて，関連文書として使われる文の意味カテゴリ，すなわち検索する対象となる研究者の研究分野に照合出来る内容も挙げられるため，これらの二つの要因を用いた分散分析を行った．

ここでは，4×4ラテン方格の横方向がフィルタリングする方法因子で，縦方向が関連文書のカテゴリ因子とする．関連文書のカテゴリ因子の水準として，32箇所のWebサイトから取って来たHTML文書を採用し，内容，またはサイト別によってそれぞれ「官文書類」，「サイエンス類」，「商社類」と「大学類」のようなカテゴリを分ける．一方，テスト1，テスト2，テスト3とテスト4では，それぞれプログラム①，②，③と④を用いて実行した．その中に，プログラム④は元の検索用のプログラムのインターフェイスのみをプログラム①，②，③のと統一したものである．

	テスト1	テスト2	テスト3	テスト4
関連文書カテゴリ1（官文書類）	A	B	C	D
関連文書カテゴリ2（サイエンス類）	B	C	D	A
関連文書カテゴリ3（商社類）	C	D	A	B
関連文書カテゴリ4（大学類）	D	A	B	C

表3. 4×4ラテン方格

6.1 実験結果：再現率

評価実験では，表示された検索結果の上位10件の内で妥当なもの数と用意した回答の割合でシステムの再現率を推測する．

変動要因	変動	自由度	分散	分散比	P-値	F境界値
関連文書カテゴリ要因	1.34	3	0.45	6.49	0.0003	2.64
テスト要因	0.32	3	0.11	1.56	0.1988	2.64
交互作用	0.85	9	0.09	1.37	0.2012	1.92
繰り返し誤差	16.50	240	0.07			

表4. 再現率の分散分析結果

表4は，繰り返しのある二元配置に従って分析した場合，各ブロックの概要と効果（平均効果）を示したものである．

表中の関連文書カテゴリ要因というのは，各関連文書カテゴリのことで，一つのカテゴリに8×2個のデータがある．テスト要因は，四つのプログラムのことを指す．

実験結果から，$F(3,63;0.05) \fallingdotseq 2.758$ であるため，観測された分散比によると，関連文書カテゴリ間に十分な有意差が認められるが，テスト間には，有意の差が認められない．すなわち，実験全体において，関連文書カテゴリが連想方式情報検索システムにおける再現率を左右するが，四つのプログラム間の差別は大きくなく，フィルタリングしても，システムの再現率を「改善した」とは言えない．

ところが，システムの再現率は，テスト要因と関係なく，入力するキーワードに大きく関連する．つまり，どのプログラムを使っても，システムにおける再現率の変化にはあまり影響を与えない．尚，再現率の変化は標準回答の絶対数との連動関係もない．再現率が低くなる理由は，関連文書の曖昧さ，つまり被験者がそれに基づいてキーワードを決める困難さと，システムの入力されたキーワードに対する漠然さという，二つのことが考えられる．

四つの関連文書カテゴリと256個の実験データに基づき，連想方式情報検索システムの平均再現率の推測値は，約0.26である．

6.2 実験結果：適合率

適合率は，検索された情報の中にユーザが求めるものがどれくらいあるかを示す割合で定義されているが，異なるプログラムでシステムの適合率を比較する時，適合情報の件数だけで，つまり適合情報の表示順を考慮しない場合，各プロセスによるフィルタリングの良さが分からないことがある．そこで，本研究では，次のような二つの方法で適合率を算出する．

①適合率の定義に従って，最後に出力されたものの上位10件のうち，標準回答に一致した件数でシステムの適合率P(com)を推測する．検索結果が10件に及ばない場合，最大表示件数で割る．

②最後に出力されたものの上位10件だけではなく，最後に現われた適合情報の順位数のうち，標準回答に一致した件数でシステムの適合率P(spe)を推測する．

それぞれの分散分析結果を，表5と表6に示す．

変動要因	変動	自由度	分散	分散比	P-値	F境界値
関連文書カテゴリ要因	2.38	3	0.79	10.21	2.36E-06	2.64
テスト要因	0.41	3	0.14	1.78	0.152502	2.64
交互作用	0.71	9	0.08	1.02	0.428542	1.92
繰り返し誤差	18.66	240	0.08			

表5. 適合率(com)の分散分析結果

変動要因	変動	自由度	分散	分散比	P-値	F境界値
関連文書カテゴリ要因	3.24	3	1.08	9.28	7.93E-06	2.64
テスト要因	0.23	3	0.08	0.66	0.579723	2.64
交互作用	1.23	9	0.14	1.17	0.315730	1.92
繰り返し誤差	27.90	240	0.12			

表6. 適合率(spe)の分散分析結果

再現率の場合と同様に，$F(3,63;0.05) \fallingdotseq 2.758$ に対して，観測された分散比には，関連文書カテゴリ間に十分な有意差が認められるが，テスト間に有意の差が認められない．しかし，これによって，テスト4以外，他の三つのフィルタリングするプロセスがシステムの適合率を「改善していない」と断言できるわけではない．なぜならば，ソートする範囲は20件に限られた場合，適合率が再現率により制限され，再現率の差が出なければ，適合率の変化は見られないこと等が考えられるからである．

しかしながら，関連文書カテゴリとテストごとに各ブロックの詳細を見ると，場合によって，三つのフィルタリング用のプロセスが異なる程度で動作していることが分かる．

まず，フィルタリングなしの検索結果より，フィルタリングした検索結果の方が，全体の平均適合率が上昇している．そのうち，最も顕著に向上したのは，関連文書カテゴリ1のテスト3とテスト4の間であった（約0.21）．さらに，この時テスト4では最小値が出たことより，表現が曖昧な官文書，およびこのような分かり難い文書こそフィルタリングする効果があることがよく分かるだろう．

次に，全体を見ると，テスト1とテスト2より，テスト3の方が安定的，且つ良い平均適合率が出ている．よって，他の方法より，コサイン尺

度モデルが，本研究におけるフィルタリングの実装に最も適切な手法であることが確認できた．

尚，tf・idfによって作られたテスト1について，P(com)とP(spe)の合計平均値の並ぶ順序が意外に変わったことから，使い方によって，tf・idfはフィルタリングの改良に特有な貢献ができると思われる．

7. おわりに

本研究によって実現されたこと，明らかにされたことをまとめると以下のようになる．

名詞高頻度語はかなり文書の意味情報を持ち，フィルタリングする方法に適用すれば，より効果的なフィルタリングが期待できる．

本研究では，フィルタリングするプロセスが主にコサイン尺度というベクトル空間型モデル手法により試作され，関連語およびキーワードで検索された結果の頻度語の高頻度語を決定するプロセスにtf・idf等の手法を投入することによって，連想方式情報検索システムの平均適合率が最大17%上昇した．

表現が曖昧な官文書，およびこのような分かり難い文書こそフィルタリングする効果があることが判明した．

連想方式情報検索システムの平均再現率が約0.26であることが分かった．この数値は，フィルタリング部の有無や適合情報の件数等と関係なく，システムの本来の仕組みに依存している．

一つの検索システムに再現率と適合率の両方を求める二つのプロセスが共存し，且つ協調的に動作することが実現したことによって，情報探索の新しいプロセスを試してみた．

今後の課題としては，プログラミング言語の考慮によって，フィルタリングにおける処理時間を短縮することや，使ったフィルタリングの手法と自然言語処理的手法との組み合せで，フィルタリングシステムのパフォーマンスをさらに向上させる等のことを挙げられる．尚，本研究，または，他の研究[10,11]で確認された名詞高頻度語に対するフィルタリングの有効な方法を発展し，クライアント側で使えるツールとして，全文検索，またはyahooやgoo等のような検索エンジンの検索結果に対応できるようにする．

参考文献

[1] Doug Oard, Jinmook Kim, Freely Available Information Filtering Systems, http://www.clis.umd.edu/dlrg/filter/software.html

[2] 奥野弘之，堀井　洋，加藤直孝，敷田幹文，國藤　進，企業ニーズを研究シーズに関連づける情報検索システムの試作，人工知能学会研究会資料，SIG-FAI-9901-36(7/9)，pp.169-174，1999

[3] 徳永健伸，情報検索と言語処理・言語と計算5，東京大学出版会，pp.199-201，1999

[4] Luhn. H. P, The automatic creation of literature abstracts, IBM Journal of Research and Development, Vol.2, No.2, pp.159-165, 1958
（邦訳）上田修一，文献の抄録の自動作成，情報学基本論文集Ⅱ，勁草書房，pp.3-16，1998

[5] Ro, Jung Soon, An evaluation of the applicability of ranking algorithm to improve the effectiveness of full-text retrieval, Journal of the American Society for Information Science, Vol.39, No.3, pp.147-160, 1988

[6] 岸田和明，情報検索の理論と技術・図書情報学シリーズ3，勁草書房，pp.73-86，1998

[7] Fredrik Kilander, Comparisons of the cosine measure and substring indexing on usenet news articles, Technical Report, Department of Computer and Systems Sciences, Stockholm University, 1996
http://www.dsv.su.se/~fk/if_Doc/T.ps.Z.

[8] 奥村　学，難波英嗣，テキスト自動要約に関する研究動向（訂正版），1999
http://www.jaist.ac.jp/~oku/okumura-j.html

[9] 佐藤理史，奥村　学，電脳文章要約術-計算機はいかにしてテキストを要約するか，情報処理，Vol.40，No.2，1999

[10] 砂山　渡，谷内田　正彦，文章要約のための特徴キーワードの発見による重要文抽出法 - 展望台システム，ことば工学研究会資料，SIG-LSE-9903-1，1999

[11] 中村勝明，國藤　進，2次元マップによる電子メールとWWWの連携ツールの試作，人工知能学会研究会資料，SIG-J-9901-16(12/18)，pp.79-84，1999