1E4-06

# 知識共有支援システム Papits における論文検索支援について

Paper Retrieval Support in Knowledge Sharing Support System Papits

藤巻 伸洋[*1] 大囿 忠親[*2] 新谷 虎松[*2]
Nobuhiro FUJIMAKI Tadachika OZONO Toramatsu SHINTANI

[*1]名古屋工業大学大学院 工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

[*2]名古屋工業大学 知能情報システム学科
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology

In this paper, we describe a mechanism for paper information retrieval. We extend existing query language for flexible paper retrieval. Our new query language allows us to decrease requery. However our query language is complex for novices to describe their information requirements. We have developed query modification system to help such users. Our modification mechanism is based on rule base mechanism which users extend query to meet the users information requirements. Our query language includes similarity based search, and provide ranking mechanism. We implement these mechanisms in the knowledge sharing support system Papits. Papits has been increasing the effectiveness in our research activity by providing noble paper information retrievals.

## 1. はじめに

近年，組織内での知識共有，知識経営に注目が集まっている．これはインターネットの普及により，デジタル情報の取得が容易になったこと，IT 技術の発展により，組織内でデジタル文書を取り扱うことが増加したことなどが主な要因となっている．このような背景から，組織内には大量のデジタル文書が蓄積されている．知識共有という観点から見ると，これら組織に蓄積された文書を知識とし，その知識を必要なメンバーが効率よく再利用できるならば，より知的生産性が向上するとされている．そこで，本研究では，知識共有支援システム Papits の開発を行っている．Papits は，組織内で取り扱うファイル，中でも論文ファイルを共有することを目的とする．

現在，Papits には，研究室のメンバーによって集められた論文が蓄積されている．これらの論文は，主に研究のサーベイなどに利用されている．その際に，Papits のユーザは，データベースに対して単純なキーワード検索を行うことができる．ここでの単純なキーワード検索とは，入力したキーワードをその論文が含むかどうかを基準とした検索である．これでは，ユーザの意図が反映されにくく，ユーザの持つ情報要求は満たされない．そこで，本研究では，ユーザが情報要求をより表現するこができるような検索質問の記述方式を提案する．しかし，提案する記述方式は，より柔軟な検索要求が表現できる反面，ユーザは複雑な検索式を作成せねばならない．そのため，システムに不慣れなユーザにとっては，意図を反映させることは難しく，慣れたユーザにも煩わしさを感じさせる要因になりうる．このような問題に対して，本研究では，検索事例を利用して，ユーザに応じた検索質問拡張を行うことで，論文検索を支援する手法を提案する．本論文では，提案手法と，その手法を実装した Papits における論文検索機構について述べる．

## 2. 関連研究

情報検索における問題点として，ユーザが要求する情報について，ある程度の知識を持っていない場合，検索質問の作成が困難となるという問題がある．このような問題に対して，検索質問作成支援に関する研究が行われている．あるキーワードで検索を行った結果から２次キーワードを生成する研究 [酒井 01] や，データマイニングによって検索質問を拡張するためのルールを導出し，検索質問作成に利用する研究 [川原 98]，視覚的に検索質問を組み立てるインターフェイスに関する研究 [砂山 00] などがある．また，ユーザモデルの推定に関しては，システムログから検索質問のパターンを生成し，それを利用することでよりよい検索システムを構築させようという研究 [鈴木 02] も行われている．

ユーザが現在注目している対象を，階層化して表示することで，情報検索を支援する研究もある．

本研究では，検索事例を用いてユーザの興味を推定し，有効な検索質問オプションが選択されるように検索質問を拡張し，ユーザが望むようなランキング結果を示すことを目的とする．

## 3. 論文検索支援

既存のシステムでは検索語を与えることで，その語を含む文書をユーザに提示するだけであった．従来の手法では，ただその検索語が含まれるだけでユーザの検索要求を満たさない文書が，ノイズとして検索結果に含まれていた．本研究では，ノイズが含まれてしまうのは，ユーザが自分の検索要求を十分に検索質問に反映させることが出来ないためであると考えた．そこで，より柔軟にユーザの検索要求を反映させることの出来るような検索質問の記述方式を提案し，それを処理することでより柔軟な検索を行う．

### 3.1 Papits における検索質問記述方式

Papits における検索質問の記述方式は，表 1 で示されているような形で，検索語 $k$ を修飾するように指定する．以下これを検索質問の修飾子と呼ぶ．検索語の修飾はネストさせることが可能なので，組み合わせによってよりユーザの意図が反映させられると考えられる．

以下に Papits で利用できる検索質問の修飾子の中での主なものについて解説する．

- domain($k$)

連絡先: 〒 466-8555 名古屋市昭和区御器所町 名古屋工業大学
知能情報システム学科新谷研究室
TEL (052)744-3153 FAX (052)735-5477
E-mail:fujimaki@ics.nitech.ac.jp

表 1: Papits で処理できる検索質問用修飾子

| | |
|---|---|
| $k$ | キーワード $k$ を含む論文を検索する. |
| domain($k$) | $k$ が一般語となる情報源を対象にする. |
| weight($k$,$n$) | $k$ の重要度を $n$ とする. |
| disambig($k$) | $k$ の表記の揺れを考慮する. |
| sim($k$) | $k$ の類似検索を行う. |
| near($k_i$,$k_j$,$n$) | 二つの $k$ が $n$ 語以内で出現する論文を検索する. |
| far($k_i$,$k_j$,$n$) | 二つの $k$ が $n$ 語以上離れて出現する論文を検索する. |
| and($k_i$,$k_j$) | 二つの $k$ を含む論文を検索する. |
| or($k_i$,$k_j$) | どちらかの $k$ を含む論文を検索する. |
| xor($k_i$,$k_j$) | どちらか一方の $k$ を含む論文を検索する. |
| not($k$) | $k$ を含む論文を検索対象から除く. |

検索語が表す分野の情報源を検索の対象として指定するための検索質問式. 指定の検索語が一般語になるような分野や, 検索語がラベルとなるような情報源に含まれる論文を検索対象とする. ここでの一般語とは,「どの文書にもその語が頻出する」ような語を指す. domain を利用することで, ユーザの意図しない論文が偶然検索語を含んでいるためにノイズとなって検索されてしまうことを防ぐ.

- disambig($k$)

  検索語の表記の揺れを考慮し, 曖昧さを解消して検索を行う. 表記の揺れとは, 同一の概念が異なる表記で表されていることを指す. 動詞の活用も表記の揺れの一つと言える. 表記の揺れを考慮することで, 検索語と完全に一致しなくとも, 同じ意味をもつ語を含む論文を検索の対象とすることができる. 実装では表記の揺れを記述した辞書と, 同義語辞書, 接辞処理を利用することで対処している.

- sim($k$)

  類似性評価知識 [大平 01] を用いて検索質問を拡張する. 類似性評価知識とは, データベースから知識発見の技術によって獲得された知識である. $R_{sim} : S \Rightarrow \langle x, y \rangle$ として表記し,「キーワード集合 $S$ とキーワード $x$ による検索質問は, キーワード集合 $S$ とキーワード $y$ による検索質問と代替可能である」ことを意味する. sim によって指定された検索語の集合に対して類似性評価知識を用いることで, 代替可能な検索質問を作成し, ユーザの表現しきれなかった意図を補うことが出来る. 類似性評価知識の獲得方法については, 文献 [大平 01] で述べられている.

- near($k_i$,$k_j$,$n$)

  二つの検索語が何文字以内で出現するかを指定し, 該当するものを検索する. 二つの語が論文中で, より接近して出現するということは, この二つの語の間で関連が強いと考えることが出来る. 語の関連は情報源の分野ごとに異なると考えられるので, この修飾子によって分野の特定が可能になると思われる.

### 3.2 Papits における検索質問支援機構

ユーザのよって作成された検索質問式は, Papits の論文検索支援機構によって処理される. 本支援機構の構成を, 図 1 に示す.

Papits では, 図 1 に示された以下の 3 つの機構によって論文検索を支援する.

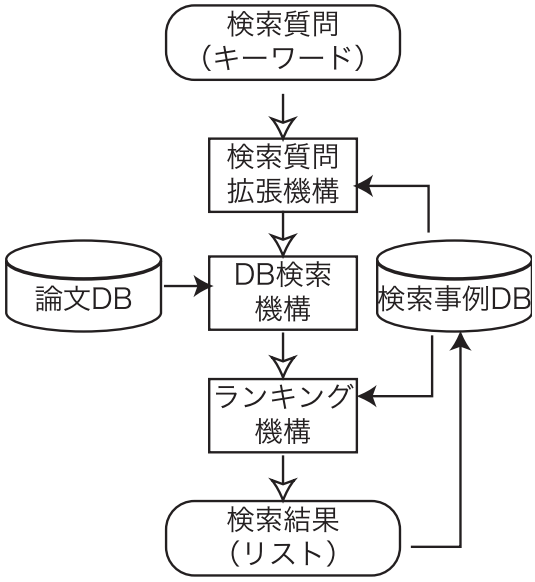


図 1: 論文検索支援機構

- ユーザが入力した検索質問を拡張する「検索質問拡張機構」
- 拡張された検索質問を用いて論文データベースを検索する「データベース検索機構」
- ユーザの興味や意図を反映し, 検索結果に順位付けを行う「ランキング機構」

検索質問拡張機構では, ユーザによって入力された検索質問を, よりユーザが望む検索結果を返すように検索質問を拡張する. データベース検索機構では, 論文データベースに対して拡張された検索質問を用いて検索を行う. ランキング機構では, データベース検索機構で得られた検索結果を, 検索要求と過去の検索事例を用いて, ユーザに適したランキングを行う.

### 3.3 検索質問拡張

ここでは, Papits における検索質問の拡張手法について述べる. Papits では, ユーザによって与えられた検索質問に含まれるキーワードに対して以下のような拡張を行う.

文献 [鈴木 02] によると, 検索語 A による検索の後で, 検索語を AB や, ABC として再検索するユーザは, 検索語 B あるいは BC によって検索結果を絞り込んでいると報告されている. そこで, 再検索の検索式を and(A,B) から and(domain(A),B) として修正する. これにより, A を一般語とする分野で, 検索語 B を含む論文を検索することが可能となる.

また，検索語 AB による検索の後で，検索語 AC として再検索するユーザは，B から C へ表記の揺れを修正していると考えられる [鈴木 02]．そこで，検索語 AB の後で検索語 AC による検索を行おうとしている場合，検索語 B と C が互いに表記の揺れの関係であるのならば，検索語 AC による検索 and(A,C) を，and(A,disambig(C)) と修正する．これによって，B,C 以外の表記の揺れも考慮した検索を行うことが出来る．

さらに過去の検索事例を用いて，検索質問を拡張する．過去に頻出する検索語は，領域を指定するための語であると考えられるので，検索語としての頻度が高いものは domain 修飾子を用いて検索式を補完する．

これらの拡張は検索質問拡張ルールとして IF-THEN 形式で表現し，処理する．それぞれを独立したルールとして表現してあるため，拡張のためのルールを追加及び，削除することは容易である．上で述べた 3 つの拡張をルールとして表現したものを以下に示す．

$$and(before(A), current(and(A, B))) \Rightarrow current(and(domain(A), B)) \quad (1)$$

$$and(and(before(and(A, B)), current(and(A, C))), ambiguous(B, C)) \Rightarrow current(and(A, disambig(C))) \quad (2)$$

$$and(frequency(A, case), current(A)) \Rightarrow current(domain(A)) \quad (3)$$

ルールを記述するために，いくつかの述語を定義した．$current(k)$ は，$k$ が現在与えられた検索語であることを示す．$before(k)$ は，$k$ が前回与えられた検索語であることを示す．$ambiguous(k_i, k_j)$ は二つの語が表記の揺れの関係にあることを示す．$frequency(k, case)$ は続く語が事例中に頻出することを示す．

検索質問拡張機構では，このような拡張ルールを用いて検索質問を拡張する．拡張された検索質問式はデータベース検索機構へ送られる．

### 3.4 データベース検索

Papits における文献の検索は，全文検索に加えて，シソーラスを用いた情報間類似性評価手法 [後藤 02] を利用する．従来のベクトル空間モデルによる手法では，多義語に対する問題が存在した．しかし，この手法では，シソーラスを利用することにより，語同士の関連を区別することで，多義語の問題を解決している．ここでは，検索語集合を一つの情報と見なし，検索語集合と，各論文との類似性を評価する．類似度が閾値を越えたものを検索結果の候補とする．シソーラスの構築については，文献 [後藤 02] で述べられている．

### 3.5 ランキング

情報源から得られた検索結果の候補に対してランキングを行う．ランキングは，なんらかの評価基準によって検索結果をスコアリングすることで処理される．評価基準は様々なものがあり，ユーザは検索質問によってランキングの手法を指定することができる．ランキングの指定方法を以下に示す．

$$(query)[: ranking[> ranking...[> ranking]]]$$

このように，ランキングキーワード ranking は検索質問の最後で指定する．複数のランキングキーワードは，不等号で示された順で優先される．以下に Papits で指定可能なランキング手法の例を挙げる．

- 被引用数によるランキング（citation）

  検索された論文の中から被引用数が多い論文を上位にランキングする．これは，検索質問によって指定された分野のオーソリティとなる論文を上位にランキングする際に有効である．

- 著者によるランキング (author)

  検索された論文の中から，最も多く登場する著者でランキングする．これは，検索質問によって指定された分野のオーソリティによって書かれた論文を上位にランキングする際に有効である．

- ダウンロード数によるランキング (download)

  組織内でダウンロードされた回数の多い論文を優先してランキングする．

- 類似性によるランキング (sim)

  検索された論文の中から，自分が持っている論文リストを情報源として比較した場合に，より類似度の高い論文を上位にランキングする．

- 時系列を考慮した類似性によるランキング (sim_time)

  自分が集めた論文の中から，より最近に集めたものに重みを付け，その上で類似度を評価し，ランキングする．最近注目している分野のサーベイや，査読などを行う場合などに有効である．

- 検索事例を考慮した類似性によるランキング (sim_case)

  Papits は，ユーザの興味を，過去の検索事例から，ユーザの興味を推定する機構を持つ．過去にユーザが検索によって発見，選択した論文には，ユーザの興味が含まれいてるものと判断する．逆に，ユーザに提示していても，選択されなかった論文は，ユーザの興味が含まれていないと判断する．Papits は，ユーザがもつ論文から，論文検索で利用したものと同様のシソーラスを構築し，そのユーザの興味のモデルとして取り扱う．これにより，ユーザと論文，ユーザとユーザの類似関係を計算することが出来る．Papits では，これらの関係を用いて検索を支援する．ユーザと類似度の高い論文は，ランキング時により上位に配置することが出来る．また，情報フィルタリングにおける協調フィルタリングの考え方と同様に，ユーザと類似度の高いユーザが持つ論文も，ランキング時に優先する．

## 4. 論文検索支援機構の実装

Papits は，WebObjects を用いて Web アプリケーションとして実装され，Web ブラウザを用いて利用する．図 2 に Papits のインターフェイスを示す．

図 2 の上部に示されているものが検索質問入力フィールド群である．各項目ごとのテキストフィールドに対し，検索質問を記述することが出来る．検索質問は，前節で述べた記述形式で記述することが出来る．発表年は，不等号や，– 記号を用いて区間指定をすることが出来る．

入力された検索質問は，検索質問拡張機構によって，拡張される．検索質問の拡張には拡張用のルールを用いる．検索質問に対して適用可能なルールがなくなるまで，ルール適用を繰り返す．すべてのルール適用が終了した時点で拡張は終了する．ルールの適用はパターンマッチングを用いて行う．

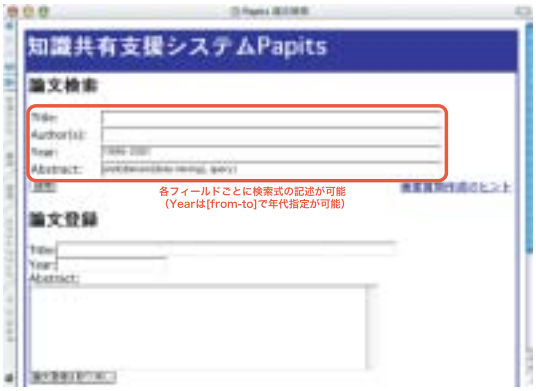


図 2: Papits による論文検索インターフェイス


図 3: Papits による論文検索結果

拡張された検索質問を用いて，データベースに対する問い合わせを行う．データベースはキーワードの全文検索以外にも，シソーラスを用いた類似性評価を行う．キーワードを含むことを前提とした検索の場合は，全文検索結果と類似性評価結果の積集合を結果とする．結果はリストとして返され，そのリストに対して，ランキングを行う．ランキングは，入力された検索質問で指定されたランキングと，ユーザの検索事例から計算される類似度に基づいて行われる．検索結果は図 3 のように出力される．

ユーザは出力結果に対して見出し語の閲覧，詳細の閲覧などを経て，必要な論文をデータベースからダウンロードする．もしくは，返された結果に満足せず，再検索を行うことも考えられる．ダウンロードをした場合，ユーザは，入力した検索質問で自らの検索要求を満たしたと考えられる．逆に，再検索を行う場合は，検索要求は満たされず，システムによる支援は失敗したと考えられる．ここでのユーザの操作は，フィードバックとして事例データベースに格納され，今後の検索に利用される．

実装したシステムの評価を行うにあたって，システムの振る舞いを観察した．その中で，$and(agent, disambig(auction))$ という検索質問に対し，$agent$ が分野をあらわす語であることが過去の事例から分かった．$disambig(auction)$ によって検索結果が膨大となってしまう可能性を含んでいるので，ここでは，検索質問拡張によって，分野をあらわす語出ることが分かっている $agent$ に対して $domain(agent)$ を適用した．これにより，ユーザの絞り込みの手間を省き，ユーザが求める情報に到達するまでの再検索の回数を減らすことになった．このことから，システムはユーザの検索を支援する振る舞いを行ったと考えられる．

## 5. おわりに

本論文では，知識共有支援システム Papits のための論文検索支援機構について述べた．ユーザが，自分の意図した情報検索が行えるように，検索要求を表現するための検索質問記述方式を提案した．

提案した記述方式は，修飾子を組み合わせることで，従来の情報検索における出現頻度に基づく検索だけでなく，検索領域の指定，表記の揺れの対応，類似文書を検索するための検索語の追加などを行うことが出来るようになった．

また，検索質問の記述方式が柔軟になることは，複雑になることにもつながるため，本論文ではユーザの意図を推測した検索質問の拡張を試みた．拡張には一般的なユーザモデルに基づく再検索時の検索質問の拡張や，過去の検索事例からユーザの興味が集中している領域を検索対象にするような拡張を行った．検索質問の拡張には書き換えルールを導入し，パターンマッチングにより適用した．ここでの問題点として，検索質問の拡張は書き換えルールの定義に大きく依存しているため，このルールを獲得する手段の発見が重要な課題となってくる．

さらに，データベース検索では，ベクトル空間モデルの多義語に対する問題点を解決した，シソーラスに基づく類似性評価手法を用いた．

検索結果に対しても，ユーザの意図した論文から順に提示するように，いくつかのランキング手法を導入した．

今後の課題としては，検索式の書き換えルールを過去の事例をもとに導出することを目指す．

## 参考文献

[酒井 01] 酒井 浩之，大竹 清敬，増山 繁：“絞り込み語提示による一検索支援手法の提案”， 言語処理学会第 7 回年次大会 発表論文集，pp. 185-188, 2001.

[川原 98] 川原稔，河野浩之，長谷川利治:“文献データベース情報検索に対するデータマイニング技術の適用”, 情報処理学会誌 Vol.39,No.4,ppV878-887, 1998.

[砂山 00] 砂山渡，大澤幸生，谷内田正彦：“ユーザの興味の構造を用いて関連検索キーを提示する検索支援インターフェイス”，人工知能学会誌，Vol.15，No.6，pp.1117-1124, 2000.

[鈴木 02] 鈴木俊輔，山名早人：” 時間間隔を用いた検索履歴のモデル化”, 情報処理学会研究報告,2002-DD-32,pp.103-110, 2002.

[大平 01] 大平峰子, 大囿忠親, 新谷虎松: “類似性評価知識を用いたレシピ検索システム mineRecipe の実装について”, 第 62 回情報処理学会全国大会論文集 (3), pp.129-130, 2001 年 03 月.

[後藤 02] 後藤将志，大囿忠親，新谷虎松: “選択的メタサーチエンジンにおけるシソーラスを用いたサーチエンジン選択手法の提案”, 人工知能学会論文誌, Vol.17, No.3, 2002 年 (掲載予定).