書籍の索引部を用いた専門家の知識に基づく検索支援システムの提案

永嶋　章弘† 　岡田龍太郎† 　本間　秀典† 　北川　高嗣†

† 筑波大学大学院システム情報工学研究科
〒 305–0006 茨城県つくば市天王台１－１－１
E-mail: †nagashima@mma.cs.tsukuba.ac.jp

あらまし　近年、コンピュータネットワーク上には多くの専門性の高い知識が蓄積されている．これらの知識を効率よく検索するためには，ユーザの要求に合致した検索支援手法の提案が不可欠であると考えられる．既存研究では，Web 上の知識全体を対象としたものが多くなされているが，専門性の高い知識を対象としたものはあまり見受けられない。本稿では，書籍の索引部から意味空間を生成し，それを意味の数学モデルに適用することで，専門家の知識に基づいて検索を支援するシステムを提案する．

キーワード　検索支援，メタデータ空間，意味の数学モデル，書籍の索引

A Search Support System Based on Domain Knowledge from the Index of a Book

Akihiro NAGASHIMA†, Ryotaro OKADA†, Hidenori HONMA†, and Takashi KITAGAWA†

† University of Tsukuba Graduate School of Systems and Information Engineering
Tenoudai 1-1-1, Tsukuba science city,Ibaraki 305–0006, Japan
E-mail: †nagashima@mma.cs.tsukuba.ac.jp

Abstract　Recently, a large amount of domain knowledge on computer networks. For using these knowledge efficiently, it's essential to realize a search support system which matches user's expectation. In this paper we propose a search support method based on domain knowledge by constructing a semantic space from the index of a book and applying it to mathematical model of meaning.

1. はじめに

近年, コンピュータネットワーク上には，膨大な量の知識が蓄積されており，これらの知識を効率よく検索するための検索支援方式が重要な研究課題となっている [8]．検索エンジンなどを用いて情報検索を行う場合の問題として，以下のような指摘がなされている．

（1） 的確に検索キーワードを選択できることは稀であり，多くのユーザが複数回の検索とキーワードの修正を強いられる傾向にある．

（2） 提示された膨大な検索結果の中から自分の求める情報を発見することは困難であり，目的を果たせずに検索を終了してしまうことが少なくない

特に（1）の問題は，単に検索実行前にユーザの意欲を殺いでしまうだけでなく，検索結果にもダイレクトに影響して（2）の問題をも助長してしまうためユーザにとってストレスとなりやすく，より深刻であると考えられる．この問題を解決するための方式として，例えば，Google サジェストのようなユーザの検索ログから検索頻度の高い検索語を提示する手法などが挙げられる．このような手法には，検索を行う前に検索しようとしている語に関連のある語を知ることができ，ユーザ自身では思いつかないような検索語を知ることができるなどの利点があるため，Web 上の知識全体を対象とした場合にはある程度有効的な手法であると考えられる．しかし，蓄積されている知識の中には，論文や特許情報のような高い専門性を持つものも多く，それらに対して検索を行う際には，このような手法の有効性は薄れてしまうことが予想される．なぜならば，専門分野における検索では，その分野の知識に基づいた専門用語を検索語として用いるのが有効であるが，その分野に関する専門的な知識を持つユーザの割合が減少するにつれて，検索ログの質が低下することが考えられるからである．また，これらの方式は“ 併せて用いられる傾向の強い ”キーワードを提示するものであり，必ずしも提示されるキーワード間に意味的な相関が認められるとは限らない．そのため，専門分野に関する知識の乏しいユーザがこれらの検索システムを利用する場合，以下のような問題に陥る可能性が想定される．

（1） 知識が乏しいため目的の情報と（意味的な）相関の強いキーワードが分からず，曖昧なキーワードで検索を開始する

（2） 検索結果を見てもどれがより目的に合っているか分からない

（3） 幾つかのページを閲覧するが，適切な修正キーワードが見つからない

（4） (1)〜(3) を繰り返し，望んだ検索結果が得られないまま検索を終了してしまう

そこで我々は，ある特定分野に関する検索を行うユーザを想定し，専門分野における検索の支援手法として，専門家の知識に基づいた検索語のサジェストを提案する．

我々はこれまでに，言葉と言葉の関係の計量による検索機構として，意味の数学モデルによる意味的連想検索を提案している．これは，単語群を文脈として解釈する機構により，言葉と言葉間の相関を文脈に応じて動的に計算することを可能とするモデルである．

意味の数学モデルを用いて各専門分野を対象とした検索を行うためには，その専門分野を表現するためのメタデータ空間を作成する必要がある．メタデータ空間とは，意味の数学モデルにおいて言葉と言葉の関係を計量するための空間を指す．我々はこれまでに，特定分野を対象とした連想検索のためのメタデータ空間生成方式を提案している．

本稿では、以上の方式を用いて専門家の知識に基づいた検索語の提示と数式の構造に着目した類似度計量による検索方式の実現に加え，Ajax を用いてページ遷移をなくす，書籍の章立てに基づいた語の提示など効率的な検索を支援するインタフェースを実装することで，科学分野を対象とした効率的な検索支援システムの実現を目指す．

これにより，その分野に関する知識が少ない場合でも，その分野に関するより専門的な知識に基づいて連想される用語を検索実行前に知ることができるようになるため，キーワード修正や検索結果の閲読にかかるストレスを軽減し，検索目的にあった情報を獲得するための支援が可能になると考えられる．

2. 意味の数学モデルの概要

本節では，言葉と言葉の関係の計量を実現する意味の数学モデルの概要を示す．詳細は，文献 [1], [2] に述べられている．

（1） **メタデータ空間 $\mathcal{MDS}$ の設定**

メタデータ空間 $\mathcal{MDS}$ と呼ばれる，検索対象となるメディアデータをベクトルで表現したデータにマッピングするための正規直交空間 (以下，$\mathcal{MDS}$) を設定する．

（2） **メディアデータのメタデータを $\mathcal{MDS}$ へ写像**

設定された $\mathcal{MDS}$ へ，メディアデータのメタデータをベクトル化し写像する．これにより，検索対象データのメタデータが同じメタデータ空間上に配置されることになり，検索対象データ間の意味的な関係を空間上での距離として計算することが可能となる．

（3） **$\mathcal{MDS}$ の部分空間（意味空間）の選択**

検索者は与える文脈を複数の単語を用いて表現する．検索者が与える単語の集合をコンテクストと呼ぶ．このコンテクストを用いて $\mathcal{MDS}$ に各コンテクストに対応するベクトルを写像する．これらのベクトルは，$\mathcal{MDS}$ において合成され，意味重心を表すベクトルが生成される．意味重心から各軸への射影値を相関とし，閾値を超えた相関値 (以下，重み) を持つ軸からなる部分空間 (以下，意味空間) が選択される．

（4） **$\mathcal{MDS}$ の部分空間（意味空間）における相関の定量化**

選択された意味空間において，メディアデータベクトルのノルムを検索語列との相関として計量する．これにより，与えられたコンテクストと各メディアデータとの相関の強さを定量化している．この意味空間における検索結果は，各メディアデータを相関の強さについてソートしたリストとして与えられる．

3. 書籍の索引部を用いたメタデータ空間生成

本節では，語とページ番号の関係が記述されている書籍の索引部を用いたメタデータ空間生成の提案方式を示す．詳細は，文献 [3] に述べられている．本方式は，全ての語を表現することができる空間を作成するのではなく，特定分野に特化した空間を生成することを目的としている．このような空間の作成を前提とすることにより，特定分野に関連する語と語の関連をより適切に表現できると考えられる．具体的には，次の流れで実現する．

（1） **初期データ行列の設定**

まず，対象とする特定分野について書かれた書籍の索引に出現する語を特徴語とみなし，索引情報から各ページ番号を用いて特徴づける．各ページを索引語を用いて特徴づけたベクトルを特徴ベクトル $\boldsymbol{p}_i$ とする．

$$\boldsymbol{p}_i = (f_{i1}, f_{i2}, \cdots, f_{in}) \tag{1}$$

ここで i はページ番号，f_{ik} は特徴語に対応したページ番号について特徴づけた値である．特徴づける f_{ik} の値は，以下のように決定される．

- 索引中で特徴語がそのページ番号を参照している場合：”1”
- 索引中で特徴語がそのページ番号を参照してない場合：”0”

文献 [2], [4], [5] のような用語辞典からデータ行列を生成する方式では，特徴づけのとりうる値を”1”，”0”，”−1”の 3 値としている．これは，用語辞典の内容から説明で「…である」などの肯定的な用法で用いられている場合は”1”,「…ではない」,「…を伴わない」などの否定的な用法で用いられている場合は”−1”と意味を読取ってデータ行列に反映させている．

本方式は索引を用いるため，索引にはキーワードとしてあらわされている語とそのページの関係しか記述されておらず，そこから，肯定の意味か否定の意味かを読み取るには，本文を逐次参照しない限り不可能である．しかしながら，語が肯定の意味に使われているか否定の意味に使われているかに関わらず，その語がそのページ番号が記載されているということは，そのページで何らかの関係を持っているということがいえる．このことから，本方式では，”1”，”0”の 2 値を用いる．

以上から，$\boldsymbol{p}_i$ を用いて，$(\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m)^{\mathrm{T}}$ とする，図 1 の

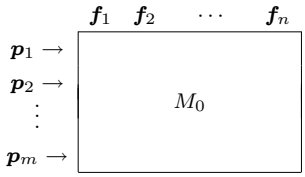

図 1　初期データ行列 M_0 によるメタデータの表現.

ような m 行 n 列の初期データ行列 M_0 を作成する.

$$M_0 = (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m)^{\mathrm{T}} \tag{2}$$

（2） 初期データ行列 M_0 の修正によるデータ行列 M の生成

(1) で作成した初期データ行列 M_0 には，ページ番号と語の関係を表す行列となっており，ページ同士の関係が反映されていない．そのため，ある概念が複数ページにわたって書かれている場合，索引に記述されているキーワードとして表される語とページ番号の関係だけでは表現しきれず，精度を悪化させる原因となりうる．初期データ行列 M_0 にページ同士の関係を反映するように修正してデータ行列 M を生成する.

一般的に，書籍には目次が付いており，目次には章，節とその題名，そしてページ番号が付与されている．章，節は，ある内容を説明するための論理的な枠であることから，内容に関係のあるページのかたまりとして捉えることができる．これらの情報を反映することにより，ページ同士の関係を反映したデータ行列 M が生成可能となる．章，節とページの関係を導入したデータ行列 M を生成することにより，章と節は各ページについて内容のかたまりによって，分けられたものあることから，内容に即した語と語の関連を導入することができる.

まず，章，節の番号を特徴語として初期データ行列 M_0 を修正，追加する．章，節番号について該当ページを全て”1”，それ以外のページを”0”と特徴づける．例えば，23 ページが 2 章 3 節に該当する場合,「2」,「2-3」を特徴語として，23 ページの「2」,「2-3」に”1”と特徴づける.

以上により，m 行 $n+\alpha$ 列のデータ行列 M を生成できる．ここで，α は章，節番号を特徴として付け加えた分である.

（3） 相関行列 $M^{\mathrm{T}}M$ からメタデータ空間生成

(2) で生成されたデータ行列 M の相関行列 $M^{\mathrm{T}}M$ を計算すると，$n+\alpha$ 行 $n+\alpha$ 列の行列となる．概要図を図 2 に示す．これは特徴語と特徴語の関係を示す行列となる．この相関行列 $M^{\mathrm{T}}M$ を固有値分解し，非ゼロ固有値に対応する固有ベクトルによってメタデータ空間を生成する．本操作の詳細は文献 [1], [2], [6] で，示されている.

これにより，語と語の関係を計量する検索のためのメタデータ空間が構成可能となる.

4. 専門家の知識に基づいた検索語のサジェスト

2. 章で述べた意味の数学モデルと 3. 章で述べた書籍の索引部を用いたメタデータ空間生成方式を用いることで専門家の知識に基づいた検索語のサジェストが実現可能となる．サジェストの手順を以下に示す.

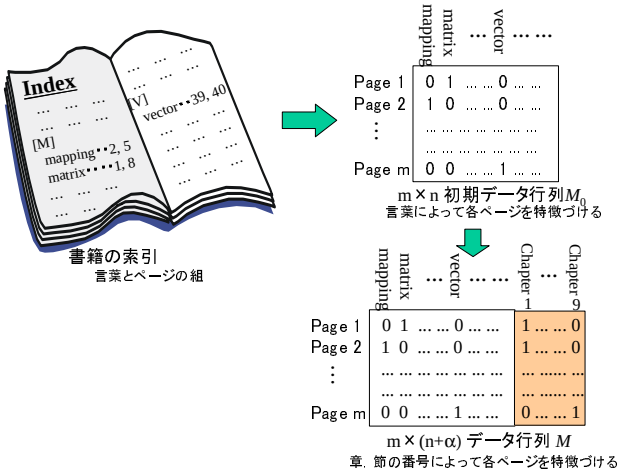

図 2　作成するメタデータの概要図.

（1） 検索語の入力

検索語を入力する．ここで入力される語は，索引語に含まれる語である必要がある.

（2） 意味的相関の計量

入力された検索語と各索引語との意味的相関を意味の数学モデルを用いて計量する.

（3） 語の提示

意味的相関を計量した結果から，相関量の大きい上位数個の語を提示する.

5. 効率的な検索支援のためのユーザインタフェース

効率的な検索を支援するためには，ユーザにとって使いやすいインタフェースが必要不可欠であると言える．そこで，本システムではユーザインタフェースに以下のような機能を実装する.

- **Ajax による画面遷移のないサジェスト**

Google サジェストのような画面遷移のない語のサジェスト機能を実装する．これにより，画面遷移によるストレスを軽減できると考えられる.

- **専門書籍の章立てを用いた検索語の提示**

本検索支援システムでは，検索を行おうとしてる専門分野についての知識をほとんど持っていないユーザを想定しており，Google サジェストのように，ユーザの入力と前方一致する語を提示する方法では，1. で述べたような状況に陥り，効果的な検索支援を行えないという問題がある．そこで本システムでは，専門書籍の章立てに基づきあらかじめいくつかの索引語を提示する方式を用いた．専門書籍の章立ては，ある内容を説明するための論理的な枠であることから，それら基づいて語を提示することは意味的に近い語を提示するという点において効果的であると考えられる．これにより，その分野の関する知識がほとんどないユーザにも効果的な検索支援ができると考えられる.

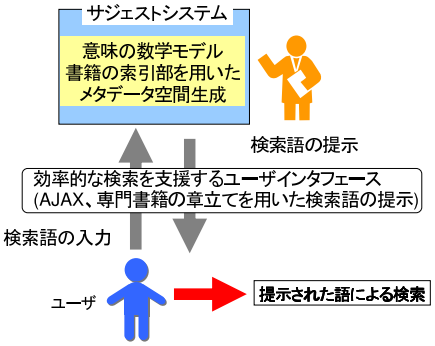

図 3　システム概要図.

6.　実装システム

本方式に基づく検索システムの実装を行った．以下に実装システムについて示す．

6.1　実 装 環 境

Web ブラウザ Firefox3 上で動作するシステムを構築した．ページ遷移をなくし，ユーザのストレスを減らすため，GUI 部には Ajax を採用した．Ajax ライブラリとして，prototype.js を利用した．サーバ側の処理には PHP を用いた．また，3. 節で述べたメタデータ空間の生成に用いる書籍には，線形計算の教科書である線形計算 [7] を用いた．

6.2　システム概要

システム概要図を図 4 に示す．検索は以下の流れで行われる．

（1）　**章立ての提示**

専門書籍に基づいた章立てを提示する．

（2）　**章の選択**

提示された章から任意のものを選択する．

（3）　**章に現れる索引語の提示**

選択された章に現れる索引語を提示する．

（4）　**専門家の知識に基づいた検索語の提示**

2. 節，3. 節，5. 節で述べた，意味の数学モデル及び書籍の索引部を用いたメタデータ空間生成方式を用いることで，専門家の知識に基づいて提示された索引語と意味的に関連のある検索語を提示する．

（5）　**検索語及び数式の選択**

提示された検索語から任意のものを選択する．

システムのスクリーンショットを図??に示す．

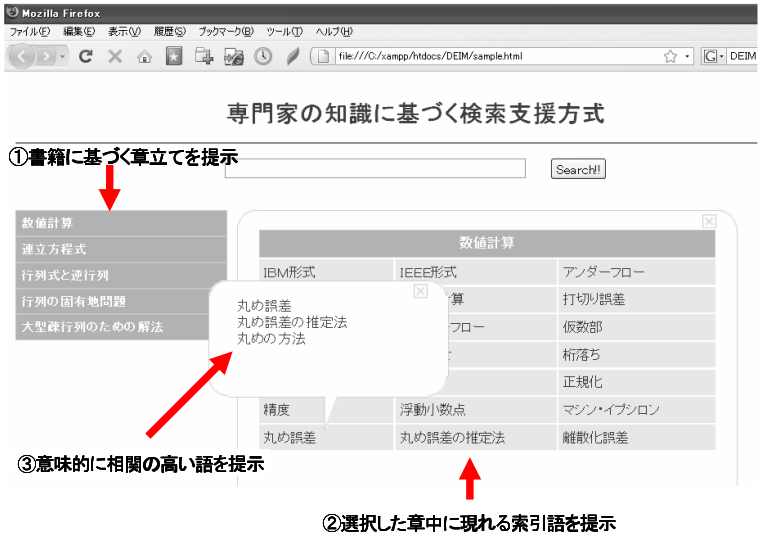

図 4　実装システムのスクリーンショット.

7.　お わ り に

本稿では，専門書籍の索引部を専門家の知識と捉えることで専門家の知識に基づいた検索語の提示について提案した．また，専門家の知識に基づいた検索語の提示及び効率的な検索支援のためのユーザインタフェースの実装を行った．

今後の課題として，複数の専門書籍のシステムへの適用，実験によるシステムの評価，ユーザインタフェースの改良，ユーザインタフェースの評価などが挙げられる．

文　　献

[1] Kitagawa, T. and Kiyoki, Y.: “The mathematical model of meaning and its application to multidatabase systems”, Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp. 130-135(1993).

[2] Kiyoki, Y., Kitagawa, T. and Hayama, T.: “A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning”, Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7 (1998).

[3] 中西崇文，岸本貞弥，櫻井鉄也，北川高嗣 “特定分野を対象とした意味的連想検索のための書籍の索引部を用いたメタデータ空間生成方式 ,” 電子情報通信学会論文誌，Vol.J88，No.4，pp.840-851,(2005).

[4] 宮川祥子，清木康，“特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式 ,” 情報処理学会論文誌：データベース，Vol.40，No.SIG5(TOD2)，pp.15-27,(1999).

[5] 河本穣，清木康，吉田尚史，藤島清太郎，相磯貞和，“医療分野ドキュメント群を対象とした意味的連想検索空間の実現方式 ,” 日本データベース学会 Letters，Vol.1，No.2，pp.12-15,(2003).

[6] 清木康，金子昌史，北川高嗣，“意味の数学モデルによる画像データベース探索方式とその学習機構，” 電子情報通信学会論文誌,D-II,Vol.J79-D-II,No. 4,pp. 509-519 (1996).

[7] 名取亮：線形計算，朝倉書店 (1993)

[8] 情報検索に対する信頼性に関する調査および結果．http://www.dl.kuis.kyoto-u.ac.jp/i-explosion/report/index.html