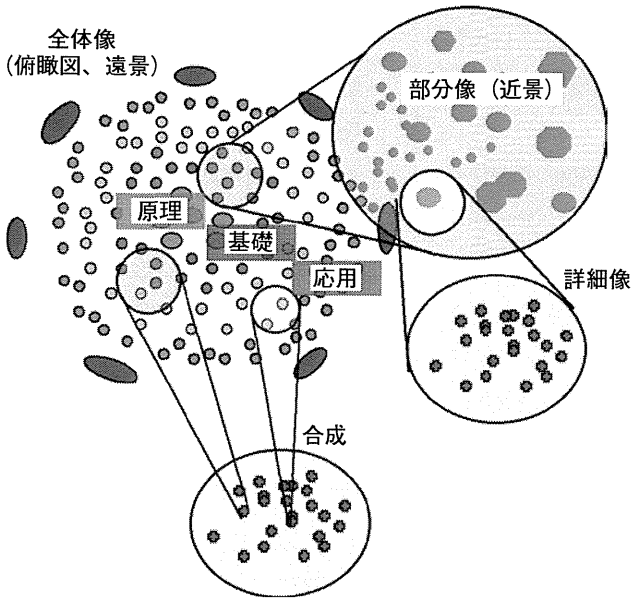

図２　知の構造化のモデル

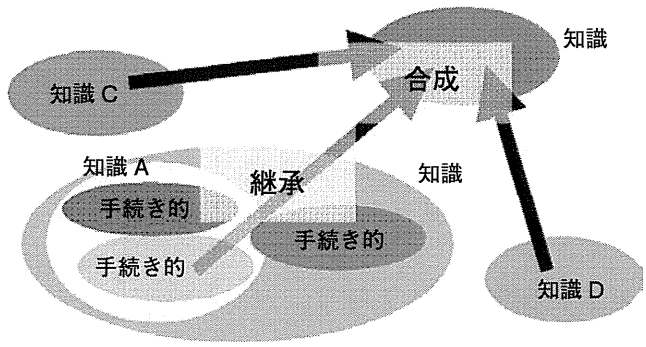

図３　知識の継承と合成

率的に活用し、価値を生み出す技術、方法論としての「知の構造化」を概説するとともに、生命科学分野への適用例をもとに、いかに可視化、構造化の技術が知の創出、活用に貢献するかについて述べる。

知の構造化の必要性

科学の拡大、深化、それにともなう分野の分化を背景に、自律分散的に創造、管理される知識 の活用に際し、問題点として以下を考える。

・情報過多、知識過多
・過度の細分化
・縦割り型(階層型)知識管理

結果として、

・知識の相互の繋がり
・知識の重複、差分
・知識の抜け

が不明瞭となっているのが現状である。

例えば、コンピュータ2000年問題、大銀行の統合のように、誰一人システムの全体像を把握していないという状況が生じる原因となる。分野を超えて知識を理解し、活用するためには、知識の全体像を明らかにすることが先決であり、総じて、「他を知る、他をわかる」ことが非常に重要となる。

知の構造化のモデル

ここでの「知の構造化」の目的は、膨大な知識を対象とし、

a) 知識間の隙間を埋める知識の発見
b) 知識間をまたぐ知識の発見

を行うことにある。一般に、すべての概念の関連は絶対的に定義できるものではないため、すべての場合において上記を厳密に区別し処理を進めることは困難であるが、すでにある例として、a)に対してはバイオインフォマティックス等、b)に対しては環境科学等がそれに当たるといえる。これを実現するための構造化モデルとして以下を考える(図2)。

1) 全体像の把握

知識の既存の関連や属性に基づく関連を抽出し、知識間の個々の関連から全体の関連を明らかにする。細分化や縦割りの弊害等により、失われがちな関連をも見つけ出すことが先決であり、オントロジー、可視化、見える化等の技術が重要な要素となる。

2) 抽象化と詳細化

膨大な量の知識の全体像を把握するためには、抽象化は必須である。抽象化された領域より、必要とする知識を選択した後、その領域の詳細化へと進めることで、必要な知識の絞り込みが容易になる。いわば、「森を見て、木を見る」操作である。

3) 合成

さまざまな知識から新たな知識を創造するためには、既存の知識を如何に再利用するかが重要である。異なる分野の知識を上記、抽象化等の操作により選択し、合成することで、より新しい知識の創出が期待される。また、創出された新たな知識を次の合成の種へとリサイクル、リファインメントをくり返すことで、知識はより成熟する(図3)。

これらの構造化、可視化、および操作が、個々人、および任意の視点によりリアルタイムに行えることが重要である。つまり、任意の視点で詳細化、抽象化の階層を上下しつつ、関連のある、もしくは関連が必要な知識を選択し、合成の要素を探すのである(図2,3)。さらには、次の瞬間

にこれら新たに創出された知識が次の合成や抽象化の対象となる。このように、知識の連続的創出と活用を促し、さらに高度な知識の再活用へと昇華させるためには、知識創出、活用の「螺旋」を形成できることが重要である。以下では、これらを支援する技術に関し、詳述する。

知の構造化システム —「MIMAサーチ」—

知の構造センターでは、上記の方法論を実践し、テキスト情報を対象とした知の構造化を支援する機構として「MIMAサーチ1」を開発し、その実用化を行ってきた。「MIMAサーチ」においては、自然言語書処理を活用することで、膨大なテキスト情報より瞬時に必要とする知識を抽出し、さらに抽出した知識間の関連性を自動で計算する。一般に、自然言語処理とは、形態素解析、構文解析、意味解析等により計算機を用いて言語の理解を行うことを指す。従来、これらを用いた仮名漢字変換、機械翻訳システム、用語(概念)抽出システム、全文検索システム等のアプリケーションが開発されており、現在では、計算機の発展により大量の言語情報を高速に処理することが可能となっている。

「MIMAサーチ」の特長は、図4に示すように、論文や、報告、アンケート等に記述されている自由文(テキスト)を自然言語処理により解析し、その統計情報に基づき、オントロジーとして重要な用語(概念)を自動的に認識・抽出することにある。

さらに、抽出したオントロジーを比較し、関連をとらえることで、文書間の意味的関連とその関連の強さを定量的に計算する。そして、それらを視覚的にとらえることができるよう、関連およびその強さをグラフモデルにより可視化する。つまり、単なる個々の文書の内容をとらえるだけではなく、文書間の意味的な関連にもとづいて全体を俯瞰し、知識を抽象化してとらえることができることを意味する。

より具体的には、「MIMAサーチ」は以下のような特徴を持つ。

- 必要とする分野全体の知識、日々創出されるリアルタイムな情報、共創的に創出される知識を含むさまざまな形態の知識群を統合し、データベースとして蓄積する。
- 上記データベースより、ある分野や領域、または分野横断的に任意の知識を抽出し、抽出された知識全体の関連を可視化する。
- 知識間の関連として、あらかじめ定義された情報、もしくは手続きにより導出される類似、包含(差分)、部分全体、因果、を含むオントロジー的関連が参照できる。
- 上記はキーワード等により指定される任意の視点を反映できる。
- 上記により指定、もしくは計算された関連をもとに、関連の強い知識同士をまとめあげる(クラスタリング/クラシフィケーション機能)。
- 上記のまとめあげを任意の抽象度で可視化する(階層的クラスタリング機能)。
- 任意の知識を選択し、また必要な知識を加えることで新たな知識を創出し、データベースに追加できる。

例えば、これらにより、複数の分野から横断的に知識を検索し、関連度指定、抽象度指定により計算されたクラスターから任意の知識を選択、さらにこれらを合成するという流れが実現可能である。

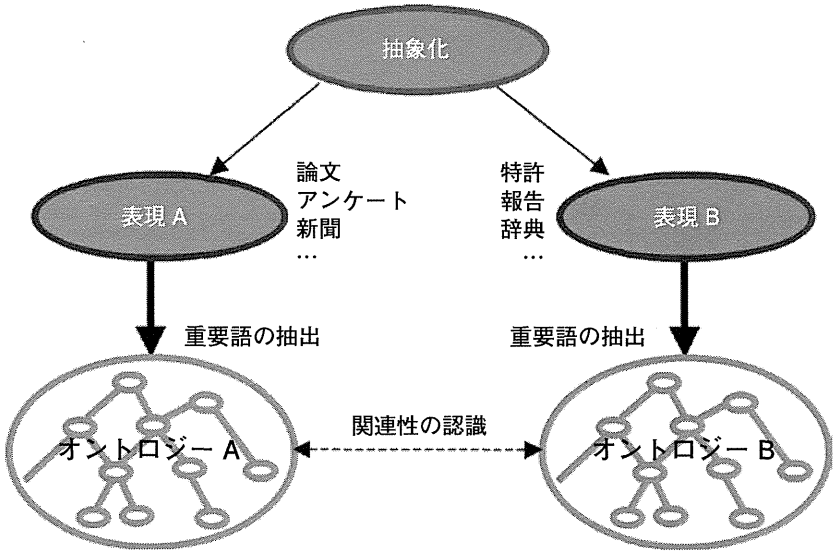

図4　特徴の抽出と関連性の認識

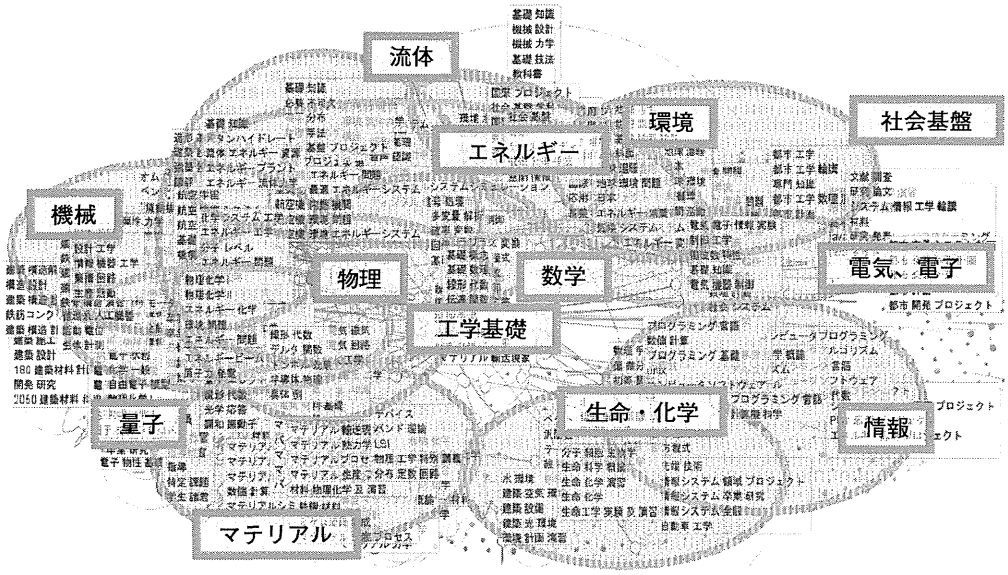

図5　工学シラバスの構造化例

例えば、図5に東京大学工学部講義シラバス(2006年度版の約900講義が対象)に対し、MIMAサーチにより処理をした例を示す。

図では、「数学」、「物理」等の原理、基礎から「情報」、「社会基盤」のような応用に至る工学知の全体像が見てとれる。また、「流体」、「エネルギー」、「環境」のような分野をまたぐ知の存在が確認できるのも特徴である。ここで注目すべきは、「環境」と「生命・化学」の隙間にある知識の"抜け"である。シラバスの充実とともに、「環境生命」のような分野の知を補うことが期待される。

なお、MIMAサーチでは、知識ソースとしてオフィスドキュメント等のファイルサーバやデータベース上にある静的知識、インターネット上の準動的知識に加え、Wikipedia等でも利用されているWikiシステムを統合することで、一般のブラウザで編集が可能な動的知識の管理にも対応している。MIMAサーチとWikiシステムとをシームレスに統合することで、検索した知識の継承や、関連する知識との合成を容易に行うことができる。また、これら継承や合成により新たに創出された知識は、即座に検索対象とすることが可能である。これにより、マルチユーザ環境では、他のユーザとも即時に知識共有が行える。

生命科学論文の構造化

生命科学分野における論文の加速的な増加は冒頭でも述べたが、爆発的に増加する論文のすべてに1人の人間が目を通すのは、すでに不可能な状況になっていることが容易に想像できる。にもかかわらず、論文の査読や発明特許の申請等においては、既知の事項との重複がないか等の、関連する分野の知識を網羅的に把握する必要がある。このような目的においても、知の構造化技術を利用することで、まずは分野全体の知識を俯瞰し、全体のなかでの位置づけをつかんだうえで、さらにその位置の詳細を確認するといった、「全体像」から「詳細像」、さらにはまた「全体像」へといったズームインとズームアウトをくり返すことで、関連する知識をより効率的に探すことが可能である。

より具体的には、まずMIMAサーチで全体を俯瞰し、意味的なまとまりのある部分に絞り込んで検索を進め、主として関連している可能性の高い論文を把握したうえで個々の関連を取り出し、検証するという詳細化のアプローチにより検索や比較の対象を絞ることが考えられる。

例えば、図6(a)(b)はそれぞれ2006年、および2007年に開催された生命科学研究ネットワーク・シンポジウムで発表された論文(それぞれ304件、324件)をMIMAサーチにより可視化したものである。

先にも述べたように、図では、内容が関連する論文がより近くになるように配置されており、よりまとまりのある論文群(クラスター)にはその内容に応じて「分子メカニズム」のような重要な用語を基に計算したトピックラベルが自動で振られている。また、さらに大きなまとまりを円で囲み、「臨床医学」のような分野名のラベルを割り当てている。年度をまたいだ恒常的なテーマが存在するのと同時に、「分子機構」や「メタボリックシンドローム」のような、それぞれの時勢に応じてテーマとなる研究が変遷していく様子が見てとれる。

また、イノベーション支援や知識創造支援の観点では、境界領域の設定のような分野を横断した関連が増加することが望ましいといえるが、図において、2006年度から2007年度への分野の変遷をみると、2006年度に比較して、2007年度には「新学術領域」と「臨床医学」や「工学」との関連が増加し、全体の繋がりがより明確になり、全体像が凝縮されていることが見てとれる。これは主に、シンポジウム等により人的交流が増加したためと推察できるが、このように、ある種の仮説検証のプロセスの一部として、知の構造化と可視化を活用することも可能である。

社会のイノベーションに対する期待が大きいなか、情報過多、知識過多により、十分活用されていない情報や知識が多く存在しているのも事実である。IT、シミュレーションなどにより、社会の効率化、自動化を目指す一方で、それを利用し、価値を生み出すのは、将来においても、やはり「人」自身である。その意味でも、膨大な情報や知識から、有用な知識やその関連を抽出し、人による知の創出、活用、価値化をいかに支援するかが「知の構造化」の本質であろう。

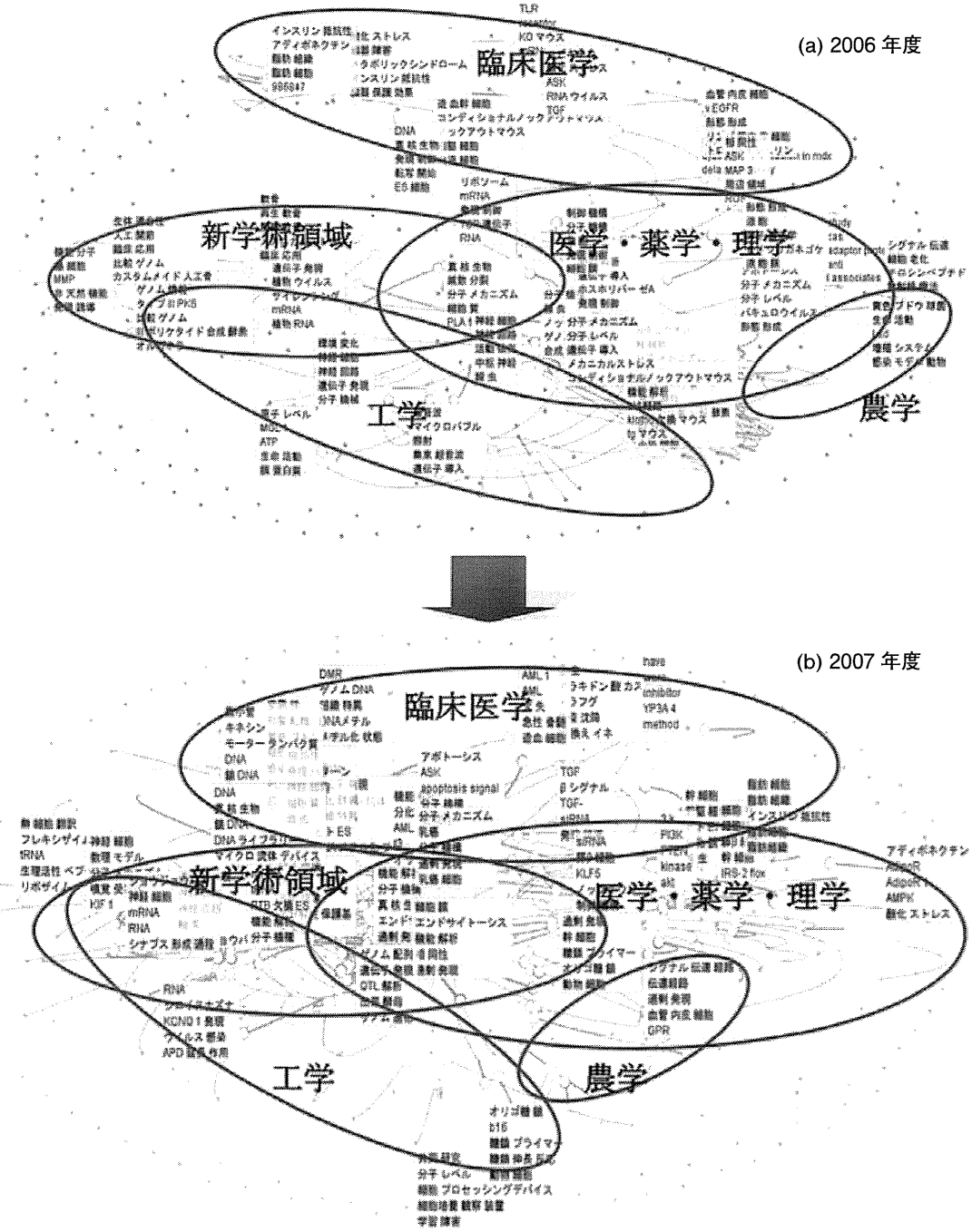

図 6　論文関連性の時間的推移

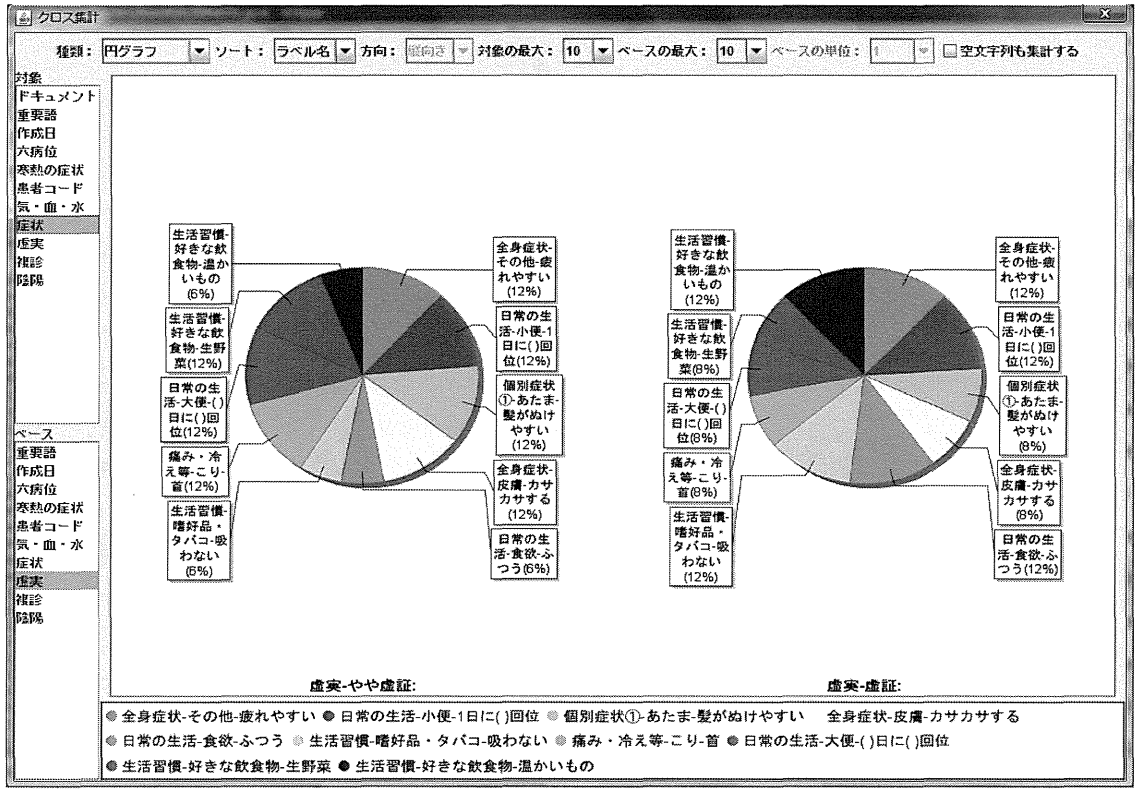

図7 「証」と問診との関連

臨床医学の構造化 —漢方診療のエビデンス創出と診断支援

本研究では、臨床医学の構造化として、漢方医学を対象とした診療のエビデンス創出、および創出されたエビデンスをもとにした診断支援システムの構築を進めている。今日では、医師の7割以上が日常診療で漢方を用いているにもかかわらず、漢方専門医以外は漢方薬の使用処方は限定されているのが現状である。これは、漢方診療が、同病異治・異病同治に示されるような、個別化診療であることが主な原因である。

また、漢方医学的診断の特徴である「証」の特定はKnow-Howのような経験知から導かれるものであり、いまだ十分に科学的・統計学的な解明が為されているとは言い難い。つまり、臨床的エビデンス(診断・治療)が得られるようになれば、一般臨床医もある一定のレベルで効果的な漢方治療ができると期待される。これを目的とし、本研究では、まず、漢方薬および鍼灸治療にともなう患者の自覚症状の推移を、外来に設置した入力端末を活用して系統的に収集し、電子カルテの医療情報とともにデータベース化を行っている。さらに、知の構造化の技術を利用することによって、治療効果の判定や漢方・鍼灸の診断「証」と症状との関連性を解析する。このようにITを活用した伝統医学の新たな臨床研究の手法を開発することで、漢方の診断と治療の科学的検証を行うことを目指している。例えば、図7には「証」(虚実)における「やや虚証」と「虚証」の診断に対する微妙な差異を統計と可視化により明確化した例であるが、本研究により、このようなKnow-Howにかかる経験的、暗黙的知識の「見える化」が行えると期待される。

実際のデータ収集においては、問診システムにて評価された患者の状態に対し、投与された薬剤や処置を問診終了後に医師が入力することとし、再診時以降は、患者の視点から評価された症状の連続的な変化が、治療経過とともに時系列で記録されている。現在、平成17年度、18年度初診患者1700名余りの診療がデータベース化されており、このデータを基にMIMAサーチによる解析を進めている。従来のこのような研究は、収集したデータの統計情報をもとに、定量的分析を行うのみであったが、さらに、MIMAサーチを利用することで可視化技術を活用した定性的分析手法との統合的解析を行っていることも本研究の大きな特徴である。

図8に、頭痛に関する患者のMIMAサーチによる分析と可視化の結果を示す。図では、グラフモデルによる患者間の関連の可視化が行われ、それぞれの点が患者を表し、患者間に引かれている線の長さと太さがそれ

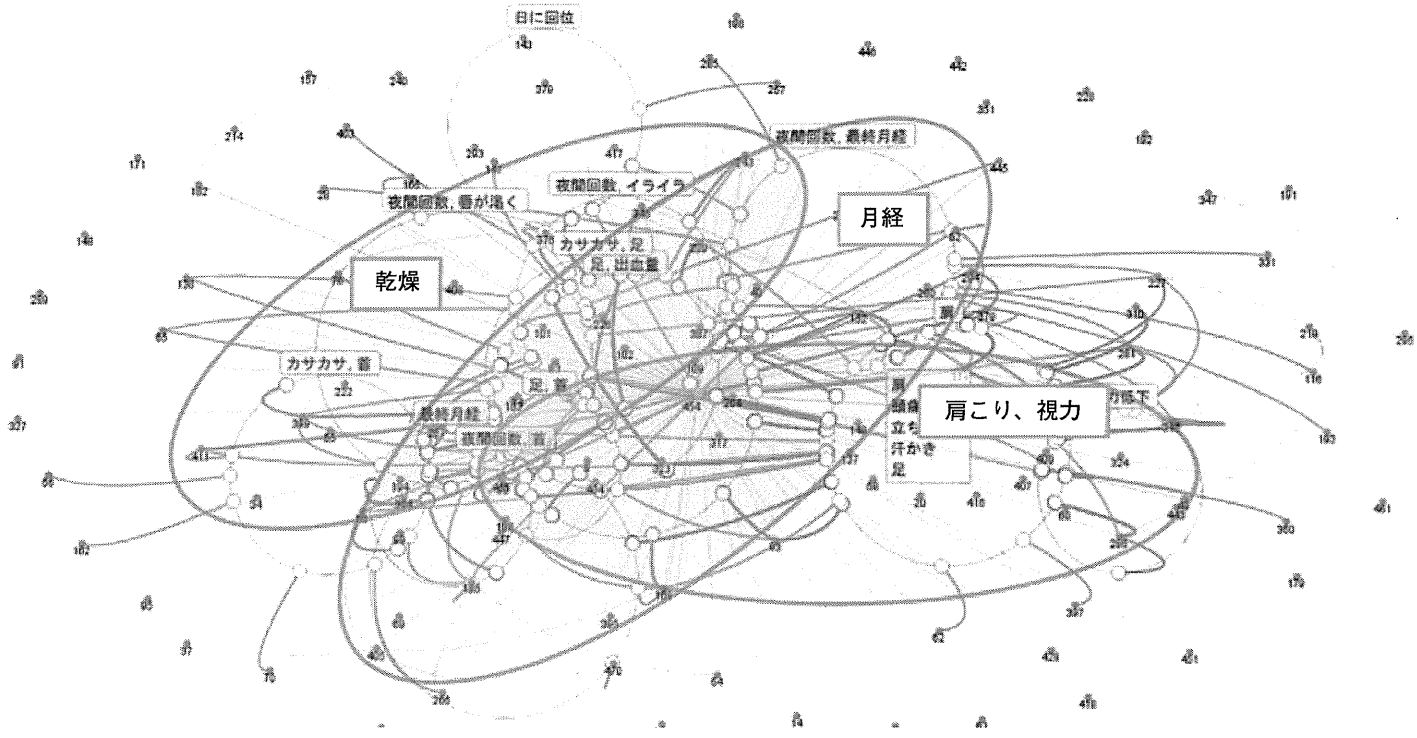

図8　頭痛に関する患者の分析結果

ら患者間の関連の強さを示す。つまり、体質や症状が似ている患者がより近くに配置されるようになる。また、図中の大きな円により関連の強い患者群(クラスター)が示され、これら円の上に示される内容が問診に表れる特徴を示している。

図からも容易に読み取れるように、頭痛に関する患者では、肩こりや視力低下、乾燥(アレルギー)、月経異常等の患者群との関連が示されている。新たな患者がどのパターンに最も近いかということや、どのような位置付けにあるか等の観察により、証の理解や、新たな治療法の開発につながるものと期待される。

溢れる情報をムダなく活かす

本稿では、分野を超えた知識の活用を目的に、可視化をはじめとする知の構造化のモデル化、また、モデル実現のためのシステム設計とその一実装である「MIMAサーチ」に関して述べた。また、実際に「MIMAサーチ」を活用した生命科学分野における知の構造化の研究を紹介した。

社会では種々雑多な情報や知識のやり取りが日々洪水の流れのように行われている。現状では、それらに1つ1つに目を通し、理解するのはほぼ不可能である。それでも、我々はそれらから意味ある情報を読み取っていかなければならない。本稿で紹介した構造化とその技術によって、日々流入する情報を構造化し、可視化すれば、それらを俯瞰することが可能になり、知識の取捨選択は大いに容易になるはずである。

リアルタイムに流れてくる情報や知識を取捨選択しつつ、価値のある知識のみを抜き取り、そこから仮説を構築し、すでに蓄積された知識を活用しながら、仮説を検証するサイクルにおいては、必要な知識を必要なタイミングで獲得できることが不可欠である。 その意味でも、知の構造化と可視化への期待は非常に大きい。

[参考文献]

1) Mima H, et al.: ACM TALIP, 5: 74-88, 2006.

Author 著者

美馬 秀樹(みま ひでき)

東京大学大学院 工学系研究科 国際工学教育推進機構 特任准教授 博士(工学)

1996年、徳島大学工学研究科システム工学専攻博士課程修了。株式会社ジャストシステム研究員、ATR音声翻訳通信研究所 研究員、英国マンチェスターメトロポリタン大学Lecturer、東京大学大学院理学系研究科 研究員、東京大学大学院工学系研究科助手を経て現職。自然言語処理、知的ユーザインタフェース、機械翻訳、知識処理、知識管理、可視化の研究と実用システムの開発に興味を持つ。平成15年度情報処理振興事業機構(IPA)未踏ソフトウェアプロジェクト 天才プログラマ/スーパクリエータ認定。平成16年 The international Daiwa Adrian prize, a triennial award for excellence in scientific collaboration between the UK and Japan, 共同受賞。2006年より英国マンチェスター大学Informatics専攻(School of Informatics)Senior honorary research fellow兼任。

An Issue-oriented Syllabus Retrieval System based on Terminology-based Syllabus Structuring and Visualization

Hideki Mima[1]
(1) School of Engineering, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
mima@t-adm.t.u-tokyo.ac.jp n

ABSTRACT

The purpose of this research was to develop an issue-oriented syllabus retrieval system that combined terminological processing, information retrieval, similarity calculation-based document clustering, and visualization.

Recently, scientific knowledge has grown explosively because of rapid advancements that have occurred in academia and society. Because of this dramatic expansion of knowledge, learners and educators sometimes struggle to comprehend the overall aspects of syllabi. In addition, learners may find it difficult to discover appropriate courses of study from syllabi because of the increasing growth of interdisciplinary studies programs. We believe that an issue-oriented syllabus structure might be more efficient because it provides clear directions for users. In this paper, we introduce an issue-oriented automatic syllabus retrieval system that integrates automatic term recognition as an issue extraction, and similarity calculation as terminology-based document clustering. We use automatically-recognized terms to represent each lecture in clustering and visualization. Retrieved syllabi are automatically classified based on their included terms or issues. The main goal of syllabus retrieval and classification is the development of an issue-oriented syllabus retrieval website that will present users with distilled knowledge in a concise form. In comparison with conventional systems, simple keyword-based syllabus retrieval is based on the assumption that our methods can provide users, and, in particular, novice users (students), with efficient lecture retrieval from an enormous number of syllabi. The system is currently in practical use for issue-oriented syllabus retrieval and clustering for syllabi for the University of Tokyo's Open Course Ware and for the School/Department of Engineering. Usability evaluations based on questionnaires used to survey over 100 students revealed that our proposed system is sufficiently efficient at syllabus retrieval.

KEYWORDS: Issue oriented, syllabus retrieval, term extraction, knowledge structuring, visualization

Proceedings of the Workshop on Speech and Language Processing Tools in Education, pages 25–36,
COLING 2012, Mumbai, December 2012.

1 Introduction

Recently, scientific knowledge has grown explosively because of rapid advancements that have occurred in academia and society.[1] This rapid expansion of knowledge has made it increasingly difficult for learners and educators to comprehend the overall aspects of syllabi. In addition, because of the rapid growth of interdisciplinary studies programs, such as energy studies and earth-environmental studies, learners have found it increasingly difficult to discover appropriate courses of study in their syllabi.

Syllabus retrieval is believed to be one of several solutions to these problems. In fact, several syllabus retrieval systems have been proposed. In general, current syllabus retrieval methods can be classified as query-oriented and/or issue-oriented. Although the query-oriented method is useful and possesses strong retrieval capabilities, it can be difficult to employ, especially for novices, because the generation of queries usually requires users to first clarify their subjects.

The issue-oriented syllabus retrieval method was developed in an attempt to provide clear directions to learners. The issue-oriented syllabus structure is believed to be more efficient for learning and education, because it requires less knowledge about subjects (Mima et al., 2006). However, this system generally requires that users classify all syllabi manually in advance. This can be a time-consuming task. Thus, we can see that it is important to develop a more efficient method for automatic syllabus structuring to accelerate syllabus classification. The advantage of this technique is based on the assumption that automatic methods will enable more efficient processing of enormous amounts of syllabi texts.

In this paper, we introduce an innovative issue-oriented automatic syllabus classification system. We integrate automatic term recognition as issue extraction, terminology-based similarity-calculation for clustering, information retrieval, and visualization. Automatically-recognized terms are used to represent each lecture (or class) in clustering. In the system, provided syllabi are automatically classified and labeled according to the included terms that were automatically extracted. The main goal of syllabus retrieval and clustering is to develop an issue-oriented syllabus retrieval website that will present distilled knowledge to users in a concise form. The advantage of this system, in comparison with conventional syllabus retrieval or classification, is based on the assumption that automatic methods can efficiently process enormous amounts of text. The system has already been put into practical use for syllabus retrieval and clustering for the University of Tokyo's Open Course Ware and for the School/Department of Engineering syllabi. Usability evaluations based on questionnaires used to survey over 100 students revealed that our proposed system is sufficiently efficient at syllabus retrieval.

In the following section of this paper, we briefly explain the process of issue-oriented syllabi retrieval. We also provide an overview of the clustering system. In Section 2, we describe our proposed syllabus retrieval and classification scheme that is based on the use of automatically-extracted terms and on a visualization technique. In Sections 3 and 4, we discuss terminological processing as a feature extraction from each syllabus for similarity calculation and

1 For example, the Medline database (http://www.ncbi.nlm.nih.gov/pubmed) currently contains over 16 million paper abstracts in the domains of molecular biology, biomedicine, and medicine. The database is growing at a rate of more than 40,000 abstracts each month.

visualization. In Section 5, we present our evaluations of data collected from questionnaires used to survey over 100 students. We relied on the collected data to analyze the usability of our proposed scheme and to confirm its feasibility and efficiency. In the final Section, we present a summary of our approach and our conclusions.

2 System Overview

The main purpose of this study was to develop an efficient issue-oriented syllabus retrieval system that would provide clear directions to learners. Our approach to this issue-oriented syllabus classification system is based on the following:

- automatic term recognition (ATR) for automatic issue extraction
- automatic term clustering (ATC) for term variation management
- terminology-based document similarity calculation to develop syllabus classification
- automatic class label inference to clarify general issues of the classes

The system architecture is modular. It integrates the following components (see, Figure 1):

- *Terminology-based issue extraction (TIE)* – A component that conducts automatic term recognition as issue extraction from syllabus texts. It includes term extraction and term variation management.
- *Syllabus retriever (SR)* – It retrieves syllabi based on selected issues that are automatically extracted by TIE. It calculates similarities between each issue and each retrieved syllabus. Currently, we have adopted tf*idf based similarity calculation.
- *Similarity Calculation Engine(s) (SCE)* – It calculates similarities between KSs provided from each KR component by the use of ontology developed by ODE to show semantic similarities between each KSs. We adopted Vector Space Model-based (VSM) similarity calculation and we used terms as features of VSM. Semantic clusters of KSs were also provided.
- *SVM-based learning (SBL)* – A component that learns how to classify syllabi by extraction of classification patterns from features that have also been extracted by TFE. It then produces classification knowledge.
- *Terminology-based syllabus classification (SBC)* – It calculates similarities between syllabi provided by the SR component by the use of terms provided from TIE to develop clusters of syllabi. We adopted Vector Space Model-based (VSM) similarity calculation.
- *Term-based label inference (TLI)* – It infers representing labels for each class developed by TSC. We currently inferred labels based on term frequency (tf) for importance and document frequency (df) for generality.
- *Syllabus class visualizer (SCV)* – It visualizes syllabi structures based on graph expression in which classes of syllabi and representing labels of classes inferred by (TLI) are automatically provided.

As shown in Figure 1 and the flows by numbers, the system extracts issues automatically from syllabi texts in advance and produces classification of lectures based on these terms or issues. Then, representing labels (i.e., class labels) are also inferred by the use of terminological information. Finally, SVC visualizes syllabi structures with respect to selected issues.

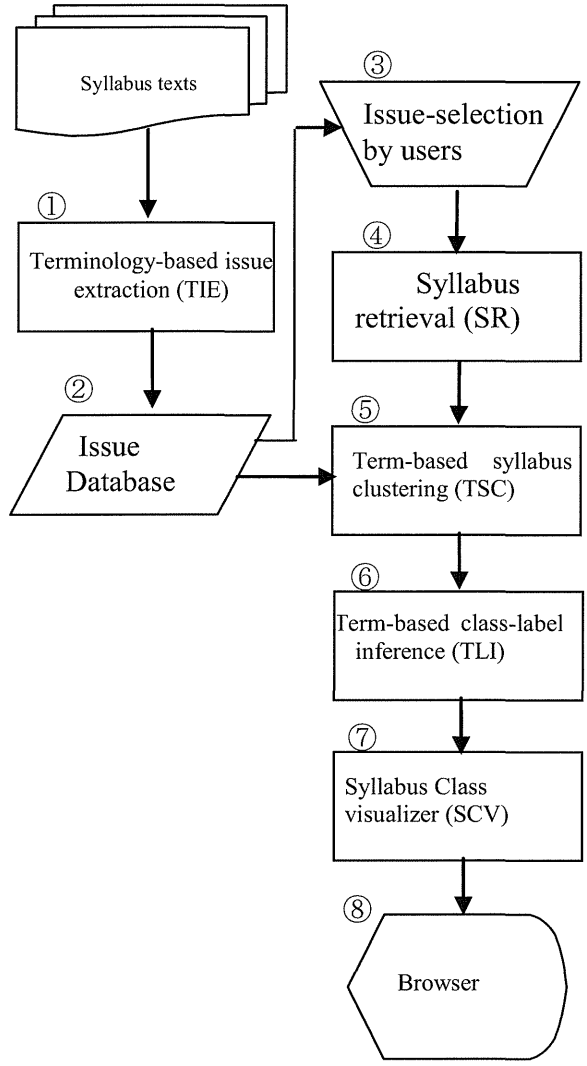

FIGURE 1 – The system diagram

3 Terminological processing as an ontology development

The lack of clear naming standards within a domain (e.g., biomedicine) makes ATR a non-trivial problem (Fukuda et al., 1998). Also, this lack of standards may typically cause many-to-many relationships between terms and concepts. In practice, two problems stem from this issue: (1) some terms may have multiple meanings (i.e., *term ambiguity*), and, conversely, (2) some terms may refer to the same concept (i.e., *term variation*). Generally, term ambiguity exerts negative effects on IE precision; term variation decreases IE recall. These problems reveal the difficulty involved in the use of simple keyword-based IE techniques. Therefore, the development of more sophisticated techniques, such as the identification of groups of different terms that refer to the same (or similar) concept(s) that could benefit from reliance on efficient and consistent ATR/ATC and term variation management methods, is needed. These methods are also important tools that can be used to organize domain-specific knowledge because terms should not be treated

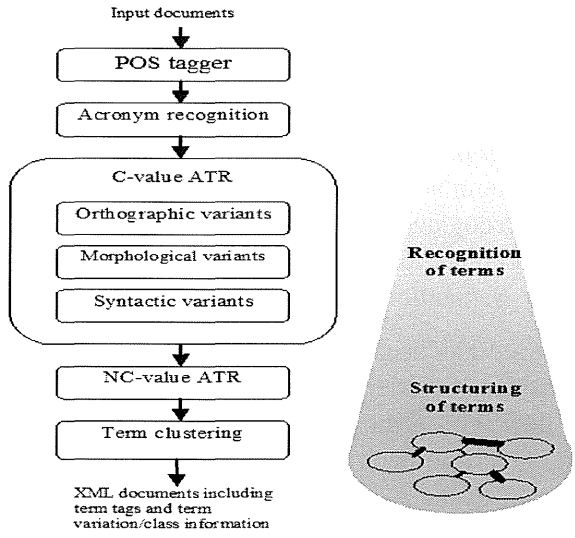

FIGURE 2 – Term recognition as issue extraction

in isolation from other terms. Rather, they should be related to one another so that relationships that exist between corresponding concepts are, at least partially, reflected in the terminology.

3.1 Term recognition

For our system, we used an ATR method based on *C/NC-value* methods (Mima et al., 2001; Mima and Ananiadou, 2001). The *C-value* method recognizes terms by combining linguistic knowledge and statistical analysis. The method extracts multi-word terms,[2] and it is not limited to a specific class of concepts. It is implemented as a two-step procedure. In the first step, term candidates are extracted by the use of a set of linguistic filters that describe general term formation patterns. In the second step, the term candidates are assigned termhood scores (referred to as C-*values*) based on a statistical measure. The measure amalgamates four numerical corpus-based characteristics of a candidate term: (1) frequency of occurrence, (2) frequency of occurrence as a substring of other candidate terms, (3) the number of candidate terms that contain the given candidate term as a substring, and (4) the number of words contained in the candidate term.

The *NC-value method* further improves the C-*value* results by considering the context of the candidate terms. The relevant context words are extracted and assigned weights based on the frequency with which they appear with top-ranked term candidates extracted by the *C-value* method. Subsequently, context factors are assigned to candidate terms according to their co-occurrence with top-ranked context words. Finally, new termhood estimations, referred to as *NC-values*, are calculated as a linear combination of the *C-values* and context factors for the respective terms. Evaluation of the *C/NC-methods* (Mima and Ananiadou, 2001) has revealed that contextual information improves term distribution in the extracted list because it places real terms closer to the top of the list.

[2] More than 85% of domain-specific terms are multi-word terms (Mima and Ananiadou, 2001).

3.2 Term variation management

Term variation and ambiguity have caused and continue to cause problems for ATR, as well as for human experts. Several methods for term variation management have been developed. For example, the BLAST system (Krauthammer et al., 2000) used approximate text string matching techniques and dictionaries to recognize spelling variations in gene and protein names. FASTR (Jacquemin, 2001) handles morphological and syntactic variations by means of meta-rules used to describe term normalization. Semantic variants are handled via WordNet.

The basic *C-value* method has been enhanced by term variation management (Mima and Ananiadou, 2001). We consider a variety of sources from which term variation problems originate. In particular, we deal with orthographical, morphological, syntactic, lexico-semantic, and pragmatic phenomena. Our approach to term variation management is based on term normalization as an integral part of the ATR process. Term variants (i.e., synonymous terms) are addressed in the initial phase of ATR when term candidates are singled out. This differs from the process that is used in other approaches (e.g., FASTR handles variants subsequently by application of transformation rules to extracted terms). Each term variant is normalized (see, Table 1, as an example) and term variants that have the same normalized form are then grouped into classes to link each term candidate to all of its variants. In this way, a list of normalized term candidate classes, rather than a list of single terms, is statistically processed. The termhood is then calculated for a whole class of term variants, rather than for each term variant separately.

Term variants	Normalized term
human cancers cancer in humans human's cancer human carcinoma	} → human cancer

TABLE 1 – Automatic term normalization

3.3 Term clustering

In addition to term recognition, term clustering is an indispensable component of the literature mining process. Because terminological opacity and polysemy are very common in molecular biology and biomedicine, term clustering is essential for the semantic integration of terms, the construction of domain ontologies, and for semantic tagging.

In our system, ATC is performed by the use of a hierarchical clustering method in which clusters are merged based on average mutual information that measures the strength of the relationships between terms (Ushioda, 1996). The system uses terms automatically recognized by the *NC-value* method and their co-occurrences as input. A dendrogram of terms is produced as output. Parallel symmetric processing is used for high-speed clustering. The calculated term cluster information is encoded and used for calculation of semantic similarities in the SCE component. More precisely, the similarity between two individual terms is determined based on their position in a dendrogram. In addition, a commonality measure is defined as the number of shared ancestors between two terms in the dendrogram. A positional measure is defined as the

sum of their distances from the root. Similarity between two terms corresponds to a ratio between commonality and positional measure.

Table 3 shows a sample of automatically-recognized terms (issues) that occur in an Engineering domain syllabus text that consists of 850 lectures (Faculty of Engineering, University of Tokyo, 2006). As we can see from the Table, reasonable and representative issues were successfully extracted by our method.

Automatically-Recognized Terms	Termhood
基礎知識 (basic knowledge)	144.55
線形代数 (linear algebra)	77.35
統計力学 (statistical mechanics)	74.00
固体物理 (solid-state physics)	67.20
ベクトル解析 (vector calculus)	65.01
偏微分方程式 (partial differential equation)	62.40
材料力学 (mechanics of materials)	62.13
環境問題 (environmental issues)	60.17

TABLE 2 – Sample of recognized issues

Further details of the methods and their evaluations can be found in Mima et al. (2001) and Mima and Ananiadou (2001).

4 The Use of Visualization to Generate Issue-oriented Syllabus Structures

In our system, the TSC, TLI, and SCV are implemented by the integration of terminology-based issue extraction from syllabi and by clustering of syllabi based on semantic similarities that are also calculated based on terms in syllabi. Graph-based visualization for the automatic generation of issue-oriented syllabus structures is also provided to help in retrieval of lectures. Figure 3 shows an example of the visualization of issue-oriented syllabus structures relevant to the issue, "environment and energy," that occurs in the engineering syllabus. To structure knowledge, the system constructs a graph in which the nodes are used to indicate relevant syllabi for the key issues selected by the user. Links among the syllabi indicate semantic similarities that are calculated by the use of terminological information developed by our TIE components. Semantic similarity is based on comparisons of terminological information extracted from each syllabus, whereas conventional similarity calculation is generally based on extracted nouns. In addition, the locations of each node are calculated and optimized when the graph is drawn. The distance between nodes depends on the closeness of their meanings. The complete algorithm of this issue-structuring method is presented below:

begin

$Q \leftarrow$ issues specified to IR

$R \leftarrow$ IR(Q) // retrieving relevant syllabi to Q and putting them into R

for every x ***in*** R ***do***

$w(Q, x) \leftarrow IRscore(Q, x)$ // *calculate IR score between Q and x*

for every y ***in*** R ***do***

if $x \neq y$ ***then***

$p \leftarrow Ont(x)$ // retrieving terminological information of x

$q \leftarrow Ont(y)$ // " y

$w(x,y) \leftarrow Sim(p,q)$ // calculate similarity using p and q

fi

end

end

Visualize graph based on every $\{w(i,j)|i=Q \text{ or } i \in R, j \in R, i \neq j\}$

end.

We generate an issue-oriented syllabus structure based on (1) cluster recognition and (2) terminology-based cluster label inference. Cluster recognition is performed by detection of groups of nodes in which every combination of included nodes is strongly linked (i.e., their similarity exceeds a threshold). Automatic cluster label inference is performed by the use of terminological information included in each cluster with respect to tf (term frequency) and df (document frequency (i.e., term generality)).

5 Evaluation

We performed a practical application of the system for syllabus retrieval for the University of Tokyo's Online Course Catalogue (UTOCC),[3] for the Open Course Ware (UT-OCW)[4] site, and for the syllabus-structuring (SS) site[5] for the School/Department of Engineering. All of these syllabi are available to the public over the Internet. The UT-OCW's course search system is designed to search the syllabi of courses posted on the UT-OCW site and on the Massachusetts Institute of Technology's OCW site (MIT-OCW). In addition, OCC and SS site's search is designed to search the syllabi of more than 9,000 lectures from all schools/departments at the University of Tokyo, and 1,600 lectures from the School/Department of Engineering at the University of Tokyo. Both systems display search results based on relationships that exist among the syllabi as a structural graphic (see, Figure 3). Based on terms that were automatically-extracted terms (issues) from the syllabi and on similarities calculated by the use of those terms, the system displays the search results in a network format that uses dots and lines. In other words,

[3] http://catalog.he.u-tokyo.ac.jp/
[4] http://ocw.u-tokyo.ac.jp/.
[5] http://ciee.t.u-tokyo.ac.jp/.

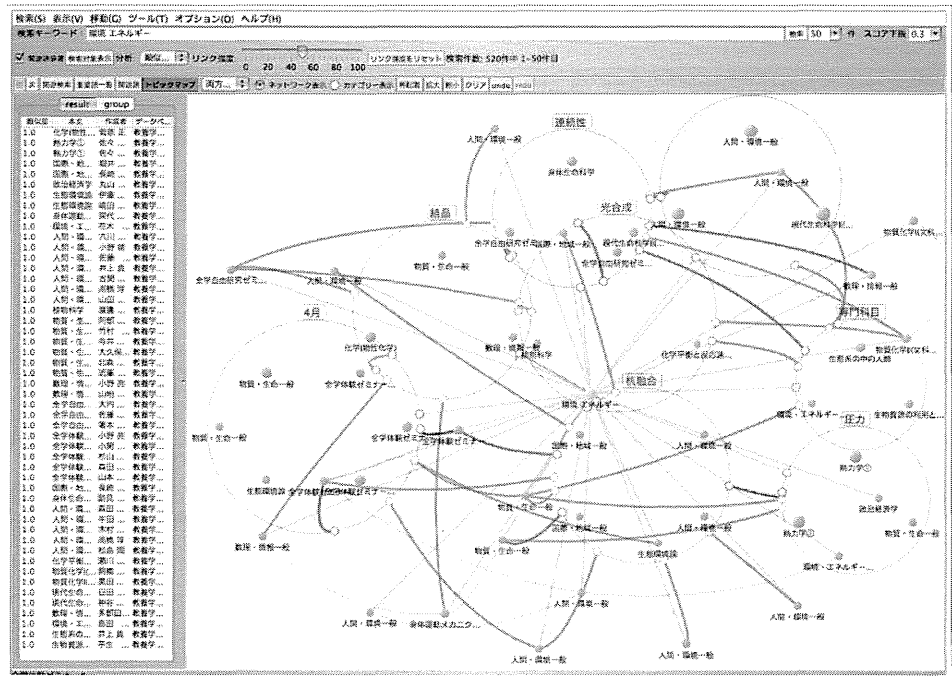

FIGURE 3 – Issue-oriented syllabus structuring: Visualization sample

the system extracts issues from the listed syllabi. It rearranges these syllabi based on semantic relationships that occur in the contents. It displays the results graphically. This differs from conventional search engines that simply list syllabi related to keywords. Because of this process, we believe users will be able to search for key information and obtain results in a minimal amount of time. In graphic displays, as mentioned previously, relevant syllabi are shown in a structural graphic that uses dots and lines with cluster circles. Stronger semantic relationships that occur in the syllabi or clusters will be located in closer proximity on the graphic. This structure will enable users to find groups of courses/lectures that are closely related to some issues. It will also help users take courses/lectures in logical order (e.g., a user can begin with fundamental mathematics and proceed to applied mathematics). Furthermore, if they consult the structural graphic display, users will be able to instinctively find relationships among syllabi drawn from other faculties or universities.

Currently, we have obtained over 2,000 hits per day, on average, from sites worldwide. We have provided more than 50,000 page views during the last three months.

We conducted a usability evaluation based on questionnaires used to survey 120 novice students. We obtained positive statements about the efficiency of syllabus retrieval by the search system from more than 70% of respondents.

Finally, we obtained 151 positive statements and 168 statements that recommended further improvement. Tables 3 and 4 demonstrate the breakdown for positive statements and the breakdown for statements that recommended further improvement, respectively. As can be seen in Table 3, we can reasonably state that our issue-oriented scheme and the related system is relatively efficient at syllabus retrieval. On the other hand, as can be seen in Table 4, we must continue to make further improvements. In particular, we must improve the visualization scheme and its scalability to link other syllabi DB and systems.

Positive statements	#
Advantage of visualization	45
Improvement in retrieval efficiency	41
Clarity of results	22
User-friendly interfaces	20
Misc.	23
Total	151

TABLE 3 – Breakdown of positive statements

Statements that recommended further improvement	#
Complexity of visualization	67
Additional linkage to other syllabi	23
Lack of clarity about relationships that exist among lectures	11
Linkage to other systems (e.g., lecture management, etc.)	13
Quality of issue extraction	10
Difficulty of operation	5
Speed of calculation	1
Misc.	38
Total	168

TABLE 4 – Statements that recommend further improvement

Conclusion

We developed an issue-oriented syllabus retrieval system that combined terminological processing, information retrieval, similarity calculation-based document clustering, and

visualization. The system provides visualizations of issue-oriented syllabus structuring during retrieval. This differs from conventional syllabus retrieval that solely provides a list of retrieved results relevant to a specific query.

We evaluated the system based on data collected from questionnaires used to survey over 100 students. Based on our results, we can reasonably state that the system provides relatively efficient syllabus retrieval.

References

Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998). *Toward information extraction: Identifying protein names from biological papers*, Proc. of PSB-98, Hawaii, pp. 3:705–716.

Mima, H., Ananiadou, S. and Matsushima, K. (2006). *Terminology-based Knowledge Mining for New Knowledge Discovery, ACM Transactions on Asian Language Information Processing* (TALIP), Vol. 5(1), pp. 74–88.

Mima, H., Ananiadou, S. and Nenadic, G. (2001). *ATRACT workbench: An automatic term recognition and clustering of terms.* In V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds.) Text, Speech and Dialogue, LNAI 2166, Springer Verlag, pp. 126–133.

Mima, H. and Ananiadou, S. (2001). *An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese, International Journal of Terminology*, Vol. 6(2), pp. 175–194.

Krauthammer, M., Rzhetsky, A., Morozov, P. and Friedman, C. (2000). *Using BLAST for identifying gene and protein names in journal articles. Gene* 259, pp. 245–252.

Jacquemin, C. (2001). *Spotting and discovering terms through NLP.* MIT Press, Cambridge MA, p. 378.

Ushioda, A. (1996). *Hierarchical clustering of words.* In Proc. of COLING '96, Copenhagen, Denmark, pp. 1159–1162.

原著論文

ニューラルネットワークと自己組織化マップを応用した川芎茶調散証の解析

竹田俊明[1)]，村松慎一[2)]

key words *senkyuchachosan, sho, neural network, self-organizing map*

【要　旨】

頭痛頻用処方のうち川芎茶調散は他の処方との関連，証が明らかでない。階層型ニューラルネットワークによる漢方処方支援システムを用いて，川芎茶調散の診断特性を検討した。藤平の頭痛頻用12処方に関する鑑別表に基づいて教師あり学習をおこなった。対象は，外来患者のうち頭痛を主訴とし川芎茶調散を処方した17名である。川芎茶調散が著効した3例のうち2例は呉茱萸湯も適応であると示された。有効の9例では釣藤散が選択されたもの3例と葛根湯が示されたもの3例があった。また，加味逍遙散，五積散，当帰四逆加呉茱萸生姜湯がそれぞれ適応とされる例があった。この手法では，各処方の推奨確率を出力するのみであるので，自己組織化マップを適用して頭痛処方全体での位置づけを試みた。その結果，川芎茶調散を含めた13処方の近縁関係を2次元座標中に表示することができた。幅広い適応のある川芎茶調散は中心付近に位置づけられ治験例の症状の解釈が妥当であることが示された。

はじめに

漢方薬治療においては，さまざまな疾病について患者の表す症状と体質の違いを考慮して処方薬（証）を決定し投与する[1)]。その診断の根拠は，傷寒論を始めとする古典的な文献や関連書籍群中に記載された症例と病態概念，師から弟子への伝承，医師自身の治療体験にもとづく病態認識と診断である。これは豊富な読解，治療経験の記憶，高度な認知活動に依存する典型的なエキスパートの作業である。

著者らは，このようなエキスパート作業の学習，獲得にはニューラルネットワーク（Neural network）が応用できることに着目し，藤平の漢方処方体系[2)]で示されている特徴判別表をインプリメント（学習過程を経て組み込む）した漢方処方診断支援システムを構築し発表してきた[3,4)]。藤平は，頭痛に対する漢方の頻用12処方についてもその特徴判別表を提示しているが，著者らが機能性頭痛（原因となる器質的疾患がなく，頭痛を主訴とするもの）

2010年2月8日受理

TAKEDA Toshiaki, MURAMATU Shin-ichi: Application of Neural network and Self-organizing map for Clarification of Senkyuchachosan-Sho

1）自治医科大学 看護学部：〒329-0498栃木県下野市薬師寺3311-159

2）自治医科大学　地域医療学センター東洋医学部門/神経内科学

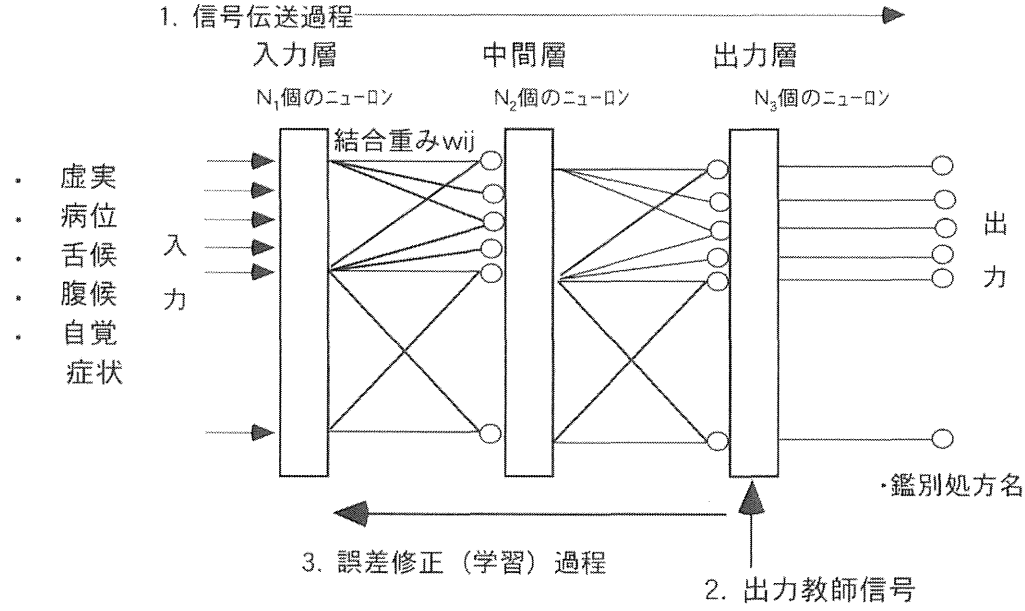

図1　誤差伝搬型ニューラルネットワーク（Back propagation method）
脈・舌・腹候，自覚症状などを入力すると，鑑別処方のなかから最適な処方が最大の数値を持つように出力される。実際に臨床的に有効であった症例から典型的な患者の症候と推奨される漢方薬の組み合わせを提示しているデータ[2]を教師信号として使用し学習を行っている。

に対して，多く適用してきた川芎茶調散[5]は示されておらず，他の頭痛処方との近縁関係が不明であり，これまでその証を明快に絞ることができなかった。

近年，片頭痛に対するトリプタン系のセロトニン受容体（5-HT$_{1B/1D}$）作動薬が開発されたことや，MRIなどの各種画像診断法の進歩にともない器質的頭痛がより確実に診断されることから，頭痛診療に際しては，国際頭痛分類に従った診断と治療が推奨されている[6]。すなわち，腫瘍・血管障害・炎症などによる二次性頭痛では原疾患に対する西洋医学的治療が優先され，一次性頭痛のうち片頭痛に対しては，トリプタン系の薬剤が第一選択とされる。このような現状において，機能性頭痛に用いて効果のある川芎茶調散の位置づけを把握し，その証がどのように表現されるか検討することは今後の漢方診療に資すると考えられた。

本研究において，ニューラルネットワークによる診断システムで川芎茶調散の診断特性を調べるとともに，位置づけを明らかにするために自己組織化マップの手法が有用であるか検討することを意図した。

Ⅰ．研究対象および研究方法

1）対象

2005年9月から2006年4月の間に，頭痛を主訴に受診した患者のうち，明らかな器質的疾患による二次性頭痛と片頭痛を除き，中等度までの頭痛が7日間以上持続している患者17名（男性5名，女性12名，平均年齢 55歳）を対象とした。川芎茶調散エキス顆粒（ツムラ TJ-124），7.5g/分3を7日間投与し，7日以内に頭痛が消失したものを著効，半減したものを有効とした[5]。

2）ニューラルネットワークによる鑑別処方の抽出

3層からなる階層型ニューラルネットワークを適用し，自己開発したプログラムを使用して教師あり学習を行った(図1)[3,4]．入力層は診断に関わる諸項目についての46ニューロンで，それぞれ藤平の漢方処方類方鑑別便覧[2]に基づき，虚実，病位，舌候（乾湿・微白苔など）・腹候（腹力・心下痞

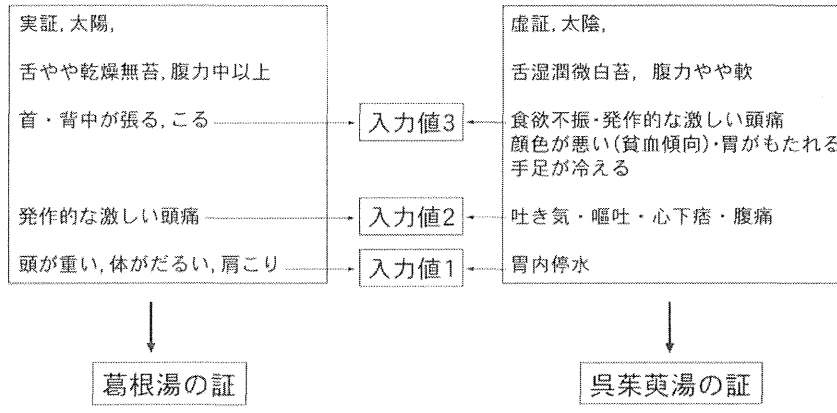

図2　入力の例
葛根湯と呉茱萸湯について，特徴的な症候と入力項目を示した。入力値3はその証に必須の症候，入力値2は全例ではないがよくみられる症候，入力値1はときにみられる症候に対する数値である。

鞭・胸脇苦満・胃内停水・瘀血など），および症状（便秘傾向・頭痛の程度・めまい・手足の冷えなど）を，それぞれの有無及び程度に対応して0，1，2，3の整数値に置き換え，入力信号とした。禁忌とされる1項目（葛根湯における「汗をかきやすい」）には-3を入力した。出力層は12ニューロンであり，処方の葛根湯，桃核承気湯，五苓散，釣藤散，加味逍遥散，三物黄芩湯，苓桂朮甘湯，半夏白朮天麻湯，五積散，当帰四逆加呉茱萸生姜湯，桂枝人参湯，呉茱萸湯が割り当てられた。中間層細胞数は前研究の結果に基づき出力細胞数と同数を採用した[3]。藤平の処方鑑別表による症状と処方の組み合わせを教師信号として，誤差逆伝搬型学習[7]を経て，収束後ネットワークを得，診断，検討に用いた（図2）。

3）自己組織化マップによる検討

自己組織化マップ（Self-organizing map, SOM）[8, 9]は，多次元（今回の解析では46の診断項目）の属性をもったデータが複数（今回の解析では川芎茶調散を含めて13の処方）あるとき，それらのデータの類似度に応じて散布図的に表現できるという理論である。2層のニューロンを用いて，教師なしで競合学習を行い，データ間の距離に応じた配置を得る。通常の2次元平面にデータを散布する場合，多次元データを2次元に圧縮して表現していることになる。

使用したデータは前項で述べた藤平の「慢性頭痛・偏頭痛への処方」[2]の特徴判別表と川芎茶調散の自験例である．特徴判別表に記載された葛根湯，桃核承気湯，五苓散，釣藤散，加味逍遥散，三物黄芩湯，苓桂朮甘湯，半夏白朮天麻湯，五積散，当帰四逆加呉茱萸生姜湯，桂枝人参湯，呉茱萸湯の12処方の症候に加えて，川芎茶調散の自験例の適用症候を入力項目の1行に組み込んだ．学習，収束を経て2次元SOMを得た。SOMの生成にはNeuralWare社のNeuralWorks Predict（SETソフトウエア（株））を用いた[10]。

II．結　果

1．川芎茶調散の効果

著効は3例（男性2，女性1），有効は9例（男性1例，女性8例），無効は5例（男性2例，女性3例）で，有効率は71％であった。著効と有効を合わせた12例についてみると，頭痛の部位は，頭全体が8例と最も多く，左側のみ，前額部，左眼窩，左側頭部が各1例あった。頭痛の性状は，頭重感，圧迫感，すっきりしないなど様々であった。当初，比較的強いギューと締め付けられるような痛みが