知能と複雑系　131－35
(2003. 1. 31)

カオス文献情報からのデータマイニングによる研究動向調査

新美　礼彦†

† 公立はこだて未来大学 システム情報科学部
〒041-8655 北海道函館市亀田中野町 116-2
E-mail: †niimi@fun.ac.jp

あらまし　本論文では、文献情報からXML+RDBシステムによる文献書誌情報データベースを構築し、構築したデータベースに対してテキストマイニング手法を適用し、タイトルや小タイトルからキーワードを抽出、そして、抽出したキーワードとと書誌情報のキーワードからキーワード解析による研究動向を調査する方法を提案する。提案した方法を実際にカオス・非線形分野に適用し、論文集からデータベース構築、主に頻度に注目したキーワード解析を行い、その結果について考察した。解析結果から、ある程度の研究分野をうかがうことができることを確認した。研究動向の解析に関しては、解析途中である。

キーワード　データマイニング、文献書誌情報、キーワード解析、研究動向調査

Research Trend Investigation from Bibliographic Database about Chaos using Data Mining Technique

Ayahiko NIIMI†

† School of Systems Information Science, Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041-8655 Japan
E-mail: †niimi@fun.ac.jp

Abstract　We proposed a method of investigating a research trend from bibliographic database using data mining techniques. This method contains, constructing a bibliographic database with XML+RDB system from document bibliography information, applying text mining techniques to the constructed database to extract keywords from paper title and chapter title, and analysing research trend using keywords from bibliographic database and extracted keywords from titles. To verify our proposed method, we applied it to the field related to chaos and nonlinear.

Key words　data mining, bibliographic database, keyword analysis, research trend investigation

1. はじめに

本論文では、文献情報から研究動向を調査する手法について検討した。

文献情報から研究動向を調査する方法は、以前より行われている。しかし最近、論文誌や学会誌、研究会報告書などがオンライン化され、ネットワーク上から手軽に検索などの利用が行えるようになった。また、パソコンの性能の向上により個人で手軽にデータ解析が行えるようになった。このような流れから、データマイニングを用いて文献書誌情報から研究動向を探ることができないか考えた。

そこで、文献データベースの構築から調査までを含めた研究動向調査の手法について検討する。研究動向調査の手順としては、次の流れを考える。まず、文献情報から文献書誌情報データベースを構築する。構築したデータベースに対してテキストマイニング手法を適用し、文献情報からキーワードを抽出する。そして、抽出したキーワードとと書誌情報のキーワードからキーワード解析による研究動向調査についてを検討する。本論文では、既存の文献データベースを利用するのではなく、文献書誌情報データベースの構築から検討することにより、より幅広い分野への適用が可能となると思われる。

本論文では、対象データとして、カオス・非線形文献データベースを取り上げ、カオス・非線形文献データベースの構築、キーワードによるカオス研究分野、研究動向の調査をおこなった。なお、タイトルに研究動向調査とあるが、本論文では、実際の研究動向調査までは行っておらず、データベース構築とデータベース解析に関する検討を行い、実際にデータベースを構築し、キーワード解析を行うところまでしか行っていない。相関

解析、年ごとの研究動向解析については、解析途中である。

2. データベース構築

文献情報からのデータマイニングを行うために、データベースを構築する。全文をデータベースに登録するのではなく、文献書誌情報のみをデータベースに登録することにした。

データベースは構築のしやすさと構築後の拡張性のため、関連データベース (RDB) と XML (eXtensible Markup Language) を組み合わせておこなった。入出力は XML を用い、データの保存に関しては RDB を用いることにより、入出力に自由度を持たせコンピュータ、人間とも可読可能な情報にでき、かつ SQL ベースによる高速なデータ検索が可能となる。

文献データベース構築に関して、自動で論文情報を検索するシステムが提案されている。[1] XML ベースの入出力を行うことにより、これらほかのシステムとの連携も容易になる。

インタフェース部分は Java の Servlet で構築する。Servlet を通じて XML と RDB とのやり取りを行うシステムとなる。システム構成の概略を図 1 に示す。

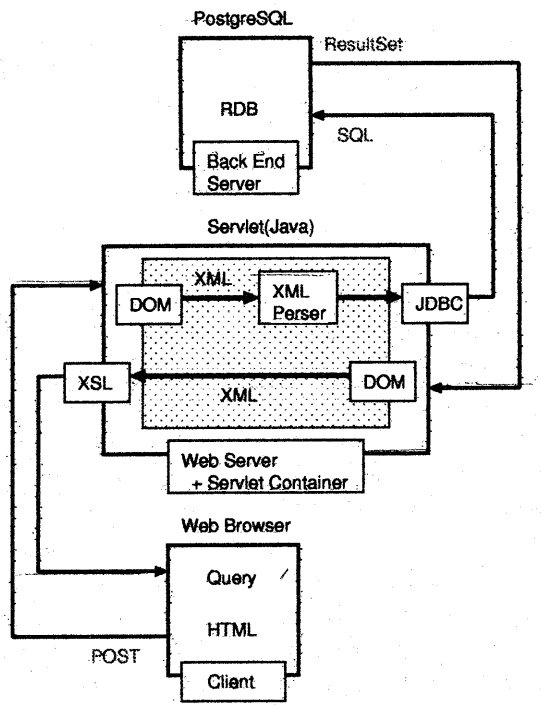

図 1 システム構成

データベースへの入力する書誌情報は、Dublin Core の 15 elements、電子情報通信学会 研究会データベース、国立情報学研究所 学会発表データベース、BiBTeX などを参考にして決定した。入出力に合わせ、RDF (Resource Description Framework) により記述した。RDF とは、メタデータなどのリソースの相互関係を記述するための規格の 1 つである。

2.1 Doblin Core

文献書誌情報を検討する際、特に Doblin Core のメタデータセットを参考にした。Dublin Core のメタデータセットはリソースを記述するメタデータとして、15 要素が提案されている。[2] (表 1 参照)

表 1 Doblin Core の 15 elements

Title	リソースの名前
Creator	リソース責任者
Subject	キーワード、
Description	要約、目次
Publisher	提供者
Contributor	協力、貢献者
Date	作成日、公開日
Type	カテゴリ、機能、分野
Format	メディアタイプ、量、サイズ
Identifier	リソースへの参照 (URI,ISBN)
Source	元リソースへの参照
Language	記述言語
Relation	関連リソースへの参照
Coverage	リソースの範囲、対象、場所
Rights	権利

Doblin Core の 15 要素は、RDF での記述が可能であるので、XML で扱いやすいという利点がある。今回は、15 要素を対象文献データに合うように一部変更して用いた。

3. データベース解析

文献書誌情報データベースからデータ解析を行う際、データマイニング手法が適用できる。データマイニングの代表的な分析手法として、頻度分析、相関分析、クラスタリングなどがある。これにより、研究分野のキーワード抽出や研究分野の広がりの調査を行うことができる。また、キーワード群を時系列的に扱うことにより、研究動向の調査として捕らえることもできる。さらに、キーワードの相関などから、用語の整理を行うことができる。

タイトルや章タイトルなどは、自然言語でかかれている。そこで、タイトルや章タイトルに対して自然言語解析の手法が使える。タイトルに対する形態素解析やそれを用いてのキーワード抽出、頻度や品詞を元にした解析、相関の高いキーワードの抽出などが考えられる。

3.1 キーワード抽出法

キーワードがつけられていない論文からキーワードを抽出するとき、タイトルや章タイトルから抽出することを考える。日本語の場合、英語のように単語の区切りが明確でないため、形態素解析などを行う必要がある。

本研究では、タイトル、章タイトルから chasen による日本語構文解析により、単語を切り出した。[3] 文章からのキーワード抽出法として、さまざまなものが提案されている。提案されているキーワード抽出法を大きく分けると、形態素解析を用いるもの、形態素解析を用いないもの、文章の構造をもとに解析するものなどがある。[4] 本論文では、主に形態素解析を用いるものを検討した。

3.1.1 形態素解析

形態素解析とは、入力文を言語学的に意味をもつ最小単位で

ある形態素に分割し、各形態素の品詞を決定するとともに、活用などの語変形化をしている形態素に対しては原形を割り当てることである。[3] 例えば、「発表会を行いたい。」という文で形態素解析を行うと、

```
発表　　発表 名詞-サ変接続
会　　会 名詞-接尾-一般
を　　を 助詞-格助詞-一般
行い　　行う 動詞-自立
たい　　たい 助動詞
。　　。 記号-句点
```

というように分析される。形態素解析で分割された単語を要素単語という。要素単語に分けることにより、頻度解析や特定品詞へのフィルタリングが行えるようになる。

3.1.2 出現頻度による抽出

形態素解析で分割された各要素単語の出現頻度を調べる。出現頻度の高い要素単語をキーワードとして抽出する。出現頻度の高い要素単語をキーワードとして抽出するため、どんな文章からも最適なキーワードを抽出しやすい手法である。しかし、助詞などのキーワードとして適切でない語を抽出する傾向があるため、抽出後のフィルタリングが重要になる。単純な頻度を使わずに、$tf \cdot idf$ を用いることもできる。これは、以下の式で定義される。

$$スコア = tf \times idf \qquad (1)$$

ただし、

tf　あるキーワードがその対象文章中に含まれる出現回数

$idf = \log(N/n)$

N　全文章数

n　そのキーワードを含むファイル数

$tf \cdot idf$ 法を用いることにより、多数の文章に多く含まれる一般的なキーワードの重要度を下げ、特定の文章中に多く含まれるキーワードの重要度をあげることができる。

3.1.3 連続名詞の抽出

情報検索の世界では名詞概念をキーワードとして抽出する傾向が強い。[5] 一般的には、形態素解析を用いて名詞を抜粋し、キーワードの抽出をおこなう。「発表会を行いたい。」という表現を形態素解析を行った結果、「発表」、[会」、「を」、「行う」、「たい」の5つの要素単語に分割される。「を（助詞）」、「行う（動詞）」、「たい（助動詞）」は、名詞ではないのでキーワードとして抽出せず、この場合「発表」、「会」といった名詞をキーワードとして抽出する。ただし「発表」、「会」といった単位では、頻度は高いが具体性が低いため、「発表会」という、長い単位で語句を抽出することにより語の具体性を上げることができる。

3.1.4 N-グラム

構文解析を行わない方法の1つとして、N-グラム法がある。N-グラムは長い文字列から部分文字列を取り出す方法で、Nには2や3などの数をとることができる。N-グラムのアルゴリズムでは1文字ずつずらしながら、連続するN文字を取り出し、取り出した文字列の共起頻度を調べ、その集合の中で共起頻度の高い語をキーワードとして抽出するというものである。[5] あらかじめ文章に品詞付けを行う必要がなく、任意の数の文字数を設定することができる。しかし、品詞付けを行わないで解析すると、単語の一部分を含んだ文字列をキーワードとして抽出する恐れがる。これを改善するために、本論文では形態素解析を行い、要素単語に分けた後で、その要素単語の連続を調べる手法も検討した。

3.1.5 相関ルール

1文中に現れる文字や単語の相関から、キーワードを抽出することが考えられる。N-グラムを用いたアルゴリズムと同様に、形態素解析を行わなくてもキーワードを抽出することが可能である。相関ルールを高速に抽出する手法として、apriori アルゴリズムがある。[6] これも、N-グラムと同様に、単語の一部分のみを抽出する可能性を減らすため、本論文では形態素解析を行った後の、要素単語間の相関ルールからキーワードを作成することを考える。

3.2 研究動向調査

文献書誌情報として入力されているキーワードやタイトルなどから抽出したキーワードを用いて、キーワードの出現頻度に注目して解析することを考える。よく使われるキーワードはその分野の中心的なキーワードとして考えることができる。

また、1つの論文中のキーワードに対して、相関ルールを考えることにより、同時に使われやすい単語を抽出することも考えられる。例えば、カオスというキーワードとニューラルネットワークというキーワードが同時に使われることが多ければ、この2つのキーワード間に不快つながりがあると考えることができる。

さらに、抽出したキーワードを用いて、研究動向を調査することを考える。まず、年ごとの頻度分析の結果から、研究会としての研究動向を調査する参考になると考えられる。また、特定のキーワードについて、年ごとの頻度分析をおこなうことにより、そのキーワードに関係する研究分野の研究動向を調査する参考になると考えられる。

また、キーワードをクラスタリングすることにより、研究分野の広がりを検討することが可能となる。年ごとのクラスタリング結果から研究動向の広がりを把握することが可能であると考えられる。

これらの解析をシステムに組み込むことにより、現在までの研究動向の把握や、新しい研究分野の開拓が容易にできるのではないかと考えている。研究を進めるための道具として、このような研究動向の解析ツールは非常に有用であると考えられる。

4. カオス文献情報データベース

解析対象のデータベースとして、カオス文献データベースを構築した。文献書誌情報データベースとして、RDB をベースとし、入出力を XML ベースにした。入力する文献書誌情報は Dublin Core を参考に決定した。データベースの入出力は Servlet を用いて実装した。データとして、電子情報通信学会非線形研究会 (信学技法) の 1959 から 2001 までを用いた。電子情報通信学会ではすでに、文献データベースがインターネッ

ト上で公開されているが、今回の解析に使うには、必要な項目が少ないこと、登録されている論文数が少ないことなどから、独自に構築した。

古い論文でアブストラクト、キーワードがついていないものが多数あったため、アブストラクトの代わりに各章のタイトルを入力することにした。解析では、タイトル、章タイトルからキーワードを抽出し、書誌情報のキーワードと合わせて解析を行った。データベースに登録する際、「はじめに」、「結論」などのあきらかにキーワードにならない章タイトルは入力していない。データ入力の際、いくつかの論文でキーワードも章タイトルもないものがあった。これに関しては、図・表見出しを章タイトルの代わりに入力した。

データベースのサイズについては表2、登録した項目については表3参照。

表2 非線形研究会データベース

電子情報通信学会 非線形研究会 (信学技法)	1959 年から 2001 年
論文数	2315
キーワード (日本語)	5881
キーワード (英語)	5953
章タイトル	14395
切り出しキーワード	9439

表3 入力した文献書誌情報

項目	対応する Doblin Core Element
タイトル (日・英)	title
キーワード (日・英)	subject
章タイトル	description
著者名 (日・英)	creator
所属 (日・英)	creator
文献番号	sdentifier
雑誌名	source
雑誌番号 (Vol, No)	source
ページ数	source
研究会名	contributor
学会名	publisher, rights
発表日	date
発表言語	language
分類 (Proceedings)	type

データマイニングをを利用して、論文を整理する方法として、バイオインフォマティクス分野をはじめてとして、様々な分野で行われている。[8] しかし、カオス・非線形分野では分野の広がりがあり、電気、数学、物理、神経系、画像、信号処理などの複数の分野にまたがっているので、解析しにくいと考えられる。本論文では、幅広い発表が行われている非線形研究会を取り上げた。この研究会では、上記の分野を全て含み、理論中心や、応用中心、メカニズム、見方、利用、式、理論中心、シミュレーション中心、実世界指向など様々な切口で捕らえることができる。

5. 解析結果と考察

タイトル、章タイトルから chasen による日本語構文解析により、単語を切り出した。抽出したキーワードと書誌情報のキーワードを合わせて、9439 のキーワードを対象に解析を行った。解析には、日本語キーワード、英語キーワード、日本語タイトルからのキーワード、英語タイトルからのキーワード、章タイトルからのキーワードのそれぞれの組合わせをつくりおこなった。

頻度解析では、上位にカオス、ニューラルネットワーク、分岐、回路モデルなどのカオス・非線形の研究分野を表しているキーワードが抽出された。(頻度解析による日本語キーワードの解析結果の一部を表4に示す。)

表4 頻度解析による結果 (上位 20)

tf.idf	keyword	*tf.idf*	keyword
418.26	カオス	96.79	カオス制御
301.57	ニューラルネットワーク	87.49	セルラーニューラルネットワーク
263.04	分岐	81.06	精度保証付き数値計算
140.88	ヒステリシス	81.06	区間解析
127.94	分岐現象	78.80	対称性
126.29	結合発振器	74.42	カオスの制御
116.36	非線形回路	71.01	非線形
105.70	同期	71.01	巡回セールスマン問題
102.77	同期現象	71.01	学習
99.80	連想記憶	67.54	制御

連続名詞の頻度解析による結果を表5に示す。ほぼ、頻度解析による結果と同じような結果になっている。

表5 連続名詞頻度解析による結果 (上位 20)

tf.idf	keyword	*tf.idf*	keyword
453.91	カオス	126.29	フラクタル
326.18	ニューラルネットワーク	116.36	非線形回路
309.33	分岐	108.60	回路
174.75	方程式	102.77	同期現象
146.50	発振器	99.80	連想記憶
143.39	制御	99.80	非線形
140.88	ヒステリシス	96.79	カオス制御
135.78	分岐現象	95.82	モデル
127.94	同期	94.75	ニューラルネット
126.29	結合発振器	94.75	ダイナミクス

N-グラムに関して、単語をベースに 1-gram,2-gram,3-gram まで解析した。ここでは、3-gram とは連続する3語の組合わせの解析である。抽出キーワードの上位には、カオス、ニューラルネットワークなど頻度分析の結果に含まれているもののほかに、方程式、モデルなどが抽出された。(3-gram までの解析による日本語キーワードの解析結果の一部を表6に示す。)

章タイトルから抽出したキーワードを用いた頻度解析の結果を表7に示す。章タイトルにつけやすい単語が抽出されているが、キーワードとして使えるものが埋もれてしまっている。論文のスタイルのようなものはうかがえるが、研究分野に直結したようなキーワードを抽出するためには、フィルタリングを工夫する必要がある。

抽出キーワードを含んだ頻度解析による結果を表8に示す。日本語、英語をまとめて解析しているので、日本語の用語を英語に言い換えたものが比較的近いスコアで抽出されている。研究分野を調べる際には、日英の言い換えなどをフィルタリング

表 6　N-グラム解析による結果（上位 20）

tf.idf	keyword	*tf.idf*	keyword
453.91	カオス	126.29	フラクタル
326.18	ニューラルネットワーク	121.91	カオスの
309.33	分岐	116.36	非線形回路
174.75	方程式	108.60	回路
146.50	発振器	102.77	同期現象
143.39	制御	99.80	連想記憶
140.88	ヒステリシス	99.80	非線形
135.78	分岐現象	96.79	カオス制御
127.94	同期	95.82	モデル
126.29	結合発振器	94.75	ニューラルネット

表 7　章タイトルからの頻度解析による結果（上位 20）

tf.idf	keyword	*tf.idf*	keyword
313.41	シミュレーション結果	136.35	基礎方程式
305.98	シミュレーション	130.90	アルゴリズム
298.14	解析結果	121.82	基本方程式
238.76	実験結果	116.29	実験方法
231.66	回路モデル	106.71	問題の記述
186.26	数値例	106.71	解析
182.38	解析方法	99.43	応用例
179.68	回路方程式	98.41	circuit model
165.71	モデル	95.72	ポアンカレ写像
137.49	数値実験	94.70	実験

表 8　抽出キーワードを含んだ頻度解析による結果（上位 20）

tf.idf	keyword	*tf.idf*	keyword
562.61	chaos	235.33	回路モデル
558.11	カオス	191.78	synchronization
379.32	ニューラルネットワーク	188.11	分岐現象
367.21	bifurcation	187.08	数値例
327.86	分岐	184.19	neural networks
316.18	シミュレーション	183.37	回路方程式
315.07	シミュレーション結果	183.18	解析方法
299.69	解析結果	178.58	モデル
294.54	neural network	174.39	hysteresis
239.91	実験結果	162.35	ヒステリシス

してまとめる必要がある。

全体的に、書誌情報からの解析のほうが、キーワード抽出処理を含んでいないぶん、きれいなキーワードが抽出された。キーワード抽出法や、その後のフィルタリングなどを検討する必要がある。特に、ほかのキーワードの一部として含まれているキーワードが多数出力されたので、これの扱いをどうするか検討する必要もある。

抽出したキーワードに関して、ほとんど同じ意味の用語 (return map, return plot etc.) や、分野による用語の違いなど単純に頻度によるフィルタリングで処理できないものがあることが確認できた。解析には、その分野の専門家との連携が不可欠であるといえる。

また、データベース中でタイトルなどに数式が多く使われていた。カオス・非線形分野では数式はキーワードとして重要であると考えられるので、今回の解析では、数式は TEX 表記を用いてデータベースに登録し、1 単語として扱った。似たような数式があったことから、数式の取り扱いについても検討する必要がある。なお、相関解析、年ごとの研究動向解析については、解析途中である。

6. おわりに

本論文では、文献情報から研究動向を調査する手法について検討し、検討した手法をカオス・非線形関係分野の文献書誌情報データベースに適用した。文献から、文献書誌情報データベースを構築した。つぎに、構築したデータベースのタイトル、章タイトルからキーワードを抽出した。抽出したキーワードと書誌情報のキーワードから主に頻度分析を用いてキーワードベースの解析を行い、それに対する考察を行った。研究動向の解析に関しては、解析途中である。

解析結果から、ある程度の研究分野をうかがうことができることが確認できた。また、解析には用語の統一が重要であるということが確認できた。

現在、データベースのクリーニングを行っており、それと並行して、多のデータマイニング手法の適用を検討している。クリーニングに関しては、用語の統一、キーワードとして不適切な用語の削除などである。キーワードレベルではなく、キーワード群や、研究分野、研究動向に関して、考察できるように実験を進めている。

文　献

[1] 高田 伸彦, 田村 武志, 大沢 一彦: XML による Web 上の論文検索システムの構築, 電子情報通信学会論文誌 D-I, Vol.J84-D-I, No.6, pp.650–657, 2001.

[2] 水田 昌明、平 博順: テキスト分類 - 学習理論の「見本市」, 情報処理 Vol.42 No.1,pp.32–37, 2001.

[3] 松本 裕治、北内 啓、山下 達雄、平野 善隆、松田 寛、浅原 正幸:日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版, 1999.

[4] 市村 由美、長谷川 隆明、渡部 勇、佐藤 光弘: テキストマイニング - 事例紹介, 人工知能学会誌 Vol.16 No.2,pp.192–200, 2001.

[5] 那須川 哲哉、河野 浩之、有村 博紀:テキストマイニング基盤技術, 人工知能学会誌 Vol.16,No.2,pp.201–211, 2001.

[6] Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules, the 20th International Conference on Very Large Databases, Santiago, Chile, Septem-

ber 1991:32pages, 1991.
[7] Doublin Core Metadata Initiative (DCMI),
http://doublincore.org/
[8] 辻井 潤一: ゲノム情報学と言語処理, 情報処理 Vol.43 No.1,
pp.29-35,2002.