修　士　論　文

Webを活用した論文サーベイ支援システムに関する研究

平成18年2月3日提出

指導教官　石塚 満 教授

東京大学大学院 情報理工学系研究科

46412　榊　剛史

内容梗概

本論文では、論文のサーベイに必要な不可欠である論文検索と論文集合の構造化という 2 つの処理に対してそれぞれ支援を行う手法を提案するまず、Web 上から論文検索を支援するために、検索クエリ拡張の手法を提案する。本手法では、Web 上の情報を用いてシソーラスを構築し、それを用いて検索クエリ拡張を行う。そして、シソーラスの適切性、およびそのシソーラスを検索クエリ拡張に用いる有効性を評価実験で示す。

次に Web 上から収集した論文をカテゴリに分類して整理するために、論文カテゴリ分類の手法を提案する。本手法では、論文のカテゴリが時系列的に変化していくことを考慮し、文書分類を文書クラスタリングの手法を組み合わせた制約付きクラスタリングという手法を提案する。評価実験により時間的に変化のある論文カテゴリへの論文を行い、本手法が既存手法より高い精度でカテゴリを再現できることにより、本手法の有効性を示す。

最後にこれら 2 つの手法の汎用性を述べ、今後の研究方針について考察する。

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景と目的

研究者が研究をする上で、既存の研究について調べ、サーベイを行う必要性が出てくる場面というのは良く遭遇するものである。我々が自分の研究に深く関係している分野について調べることで、どのような手法を用いれば自分の研究の目的を達せられるか、ということを知ることができる。また自分の研究が当該分野においてどのような位置に位置づけられるのか、ということや、他の研究との差異を知ることで自分の研究の強みを知ることができるだろう。また自らの研究と関係ない分野をサーベイすることで、新たな研究の方向性を模索したり、そこから問題解決のヒントを得ることもできる。このように研究をする上で、論文を収集し、サーベイすることは非常に重要な作業であるといえる。

以前はこのようなサーベイ作業は非常に骨の折れる仕事であった。なぜなら、まず当該分野にどのような論文があるのか、またどこにいけばその論文が手にはいるのか、と言った情報を検索可能なシステムがあまり整備されておらず、自らが会員となっている学会の学会誌や発売されている書籍等から知るしかなかった。また、論文を手に入れるのも直接大学の図書館に出向いたり、郵送で送ってもらう等の手段しかなかった。まして最新の論文を手に入れることなどは非常に困難であった。

しかし、近年 Web 上の文書の爆発的な増加や情報検索の技術の飛躍的な更新により、我々は容易に論文を手に入れられるようになった。Yahoo や Google といった一般的な検索エンジン以外にも、GoogleScholar や CiteSeer といった Web 上に存在している論文を検索するための論文専門の検索エンジンも登場している。そして、研究者自身が自分の論文を Web にアップするだけでなく、学会や研究所なども積極的に論文を Web 上に公開しており、多くの研究者が閲覧できるようになってきている。特に情報工学の分野はその傾向が著しく、最近では出版されるよりも早く Web 上から新規論文を取得できるようになってきている。このように、現在は、以前と比べて我々は非常に多くの論文にアクセスできるようになっている。

しかし、大量の論文にアクセス可能であることは、同時に多くの問題を生みだしている。一つは目的とする論文の検索の困難さである。現在では、似通ったタイトルの論文が多数存在しており、検索エンジンなどを用いてもなかなか目的としている論文が探し出せないと言う問題がある。例えば「Detecting Community Structure in Networks」というタイトルの論文を探したいと思っても、「Fast algorithm for

detecting community structure in networks」や「An Agent-Based Algorithm for Detection Community Structure in Networks」など、類似したタイトルの論文が多数存在している。

二つ目は、ある分野に関する論文の収集の困難さである。例えば「シソーラス」と検索しても、その語をタイトルや本文に含んでいる全ての論文がヒットしてしまうためにどの論文が自分の探し求めている分野に該当する論文であるかどうかはやはり人間が目を通して判断する必要がある。三つ目は、重要な論文の発見である。例えば以前であれば、情報が少なくそれが整理されていたために、重要な論文を容易に知ることができたが、現在では、手に入る論文が非常に多いため、どの論文が当該分野において重要であるか、鍵となっているかなどを判断することが非常に難しくなっている。論文検索システムでは、被引用数をもとに論文の重要度などを表示しているが、被引用数が多い論文が必ずしも重要な論文であるわけではなく、決定性にかけるものとなっている。

このように大量の論文へのアクセスが可能になり、論文は手に入れやすくなった。しかし同時に、その大量さ故に整理されていない情報が増え結果的に重要な論文、目的とする論文を探す作業が困難となっている。

そのような背景の中で、近年論文のサーベイ作業を支援するシステムや電子図書館化システムの手法などが提案されている。前述の GoogleScholar や CiteSeer もそうであるし、研究者が作り運営している PRESRI や国立情報学研究所などが運営している CiNII なども当てはまる。本論文では、これらのようなシステムをサーベイ支援システムと呼ぶ。サーベイ支援支援システムは、Web からの論文の検索や、論文の重要性の判断、内容的に類似した文献の参照など、サーベイをより補助するような機能をサポートしており、今後ますます発展していく分野であると考えられる。

サーベイ支援システムがサポートしている機能は大きく分けて以下の 2 つである。

1. 論文検索機能

2. 論文集合の整理・分類機能

論文検索機能は、ほとんどのサーベイ支援システムがこの機能をサポートしている。研究所や学会のシステムでは、所有している論文の書誌情報を電子化し、検索インデックスとして利用可能にしている。また、Web から論文をクローリングし、収集しているシステムでは、書誌情報を自動的に取り出し、インデックス化する技術が整備されている。しかし、自分の目的としている分野の論文を探し当てるのは難しく、検索のクエリ変えて何度も検索を行う必要がある。

もう一つは、論文集合の整理・分類機能である。これは多くのシステムでサポートされているわけではないが、論文間の引用・被引用関係からその論文の重要性を判断する機能や、内容的な類似度から関連論文を参照可能にしているシステムもある。しかし、こちらはまだ未整備であり、精度もそれほど高いとは言えない。

本論文では、上記2つの機能を考慮した新たな論文サーベイ支援システムを提案する。

論文検索機能については、検索クエリの選定の困難さに着目し、Web上の情報を用いて検索クエリの拡張を行うことで、より効率的な論文検索を目指す。

論文集合の整理・分類機能については、論文集合を新たな手法用いて分析することで、論文集合内での分野の時系列的な変化を捉えることを目標とする。

1.2 本論文の構成

本論文では、検索クエリ拡張による論文検索と論文集合の構造化の2つのトピックについて扱っている。

第2章では、既存の論文サーベイ支援システムについて紹介し、それらではそれほど効率のよいサーベイ支援ができないことを指摘する。そして、本論文で提案するサーベイ支援システムについて述べる。

第3章では、従来の検索クエリ拡張の手法を紹介し、論文という専門用語の多い分野に適用する困難さを指摘する。そして、それを解決するために検索エンジンを効率的に利用した手法を提案し、その評価を行う。

第4章では、論文のカテゴリ変化を捉えることに焦点を当て、その目的に従来の文書整理手法の適用可能性を考える。そして、従来手法の特徴を併せ持った提案手法を考案し、それカテゴリ分類の評価を行う。

第5章では、本研究での提案をまとめ、さらに今後の研究方針について考察する。

第2章　サーベイ支援システムの概要

サーベイ支援システムとは、研究者がある分野について行うサーベイ作業を支援するシステムである。サーベイ支援、といってもその意味する範囲は幅広く、単純にあるキーワードを含む論文を表示から、タイトルや著者、発行年などの項目別検索、引用・被引用文献の参照、類似文献の検索まで、その機能は実に多岐に渡っている。以下では、既存のサーベイ支援システムを紹介しながら、その問題点を述べ、さらに提案システムの概要について説明する。

2.1　既存のサーベイ支援システムとその問題点

既存のサーベイ支援システムは論文の収集方法によって 2 つに分けることができる。1 つは、学会や出版社、研究所等が所有している論文を整理し、タイトルや著者名、本文などをインデックスとして検索可能にしたシステムである。ACM Protal[1]や arXiv[2]、CiNII[3]などが有名である。もう 1 つは、研究者や企業が Web 上をクローリングすることで論文を収集し、それらを整理し、タイトルや著者名、本文などをインデックスとして検索可能にしたシステムである。Google Scholar[4]や CiteSeer[27]、PRESRI[58] などが有名である。

本節では、これらの既存のサーベイ支援ステムを紹介し、それらの特徴についてまとめる。

2.1.1　大学・学会・研究所が提供するサービス

大学や学会、研究所が提供するサービスは、それぞれの機関が所有している論文にインデックスをつけ、整理したシステムである。

CiNii

NII（国立情報学研究所) によって提供されるシステムである (図 2.1)。日本語の論

[1]http://portal.acm.org/portal.cfm
[2]http://arxiv.org
[3]http://ci.nii.ac.jp/cinii/servlet/CiNiiTop#
[4]http://scholar.google.com

文が検索可能なシステムとして・は最も大規模なものである。

図 2.1: CiNII の画面

- 大規模なデータベース
  2006 年 2 月 1 日時点で、約 10,000,000 件＝書誌情報 約 3,000,000 件＋引用文献情報 約 7,000,000 件。
- 人手によるタグ付け
  人手によるタグ付けが行われているため、高精度な書誌情報による検索が可能
- 引用・被引用文献の表示
  検索した論文が引用している文献・引用されている文献へのリンクが表示される。
- 本文の表示
  NII の電子図書館 NACSIS-ELS と連動しており、そこに存在する論文は本文の表示が可能。ただし、これは有料サービスである。
- 所蔵場所の表示
  同じく NII の目録検索サービス Webcat と連動しており、その論文を所蔵している図書館が分かる。

このようにCiNiiは、本文閲覧に制限があるものの、そのデータベースの大きさと機能・使いやすさから、日本語論文を検索するシステムとしては非常に優れていると言える。

ACM Portal

ACMは世界最古・最大の教育および科学コンピューティング協会である。数多くの論文誌が参照可能である。

- 大規模なデータベース2006年2月1日時点で、コンピュータに関する文献が約750,000件所有。
- 人手によるタグ付け
  人手によるタグ付けが行われているため、高精度な書誌情報による検索が可能。
- 引用・被引用文献の表示
  検索した論文が引用している文献・引用されている文献へのリンクが表示される。
- 類似文献の表示
  内容的に類似した文献が表示される。
- 本文の表示
  多くの論文に対して、本文も表示することができる。ただし有料サービス。
- 詳細な分野分類

これらのシステムでは、人手により書誌情報や参考文献の電子化が行われているため、情報の確実性が非常に高い。そのため、タイトルや著者名など書誌情報に基づいた論文検索や、引用・被引用文献の参照を高い精度で行うことができる。また、論文を所有しているため、その論文が手に入れられること、特に最新の論文まで手に入れられることは長所の一つである。その反面、論文本文の解析や、論文同士の関係性の解析などは行っていないため、あるキーワードに強く関連している論文やある分野で重要な論文などを検索することは難しくなっている。また、あくまでその組織が所有している論文が対象であるため、検索範囲が限られてしまうという欠点もある。

2.1.2 研究者による論文検索システム

Cite Seer

CiteSeer[27] はNEC Research InstuituteのS.Lawrenceらによって作られたComputer Science に特化した論文検索システムである。WWW 上に存在する英語論文を Crawring することで収集して構築された。このシステムの大きな特徴は以下の４つに分けられる。

- 大規模なデータベース
  2004 年 11 月 1 日の時点で、716797 の論文から検索可能

- 引用・被引用文献の表示
  図 2.2 のように検索した論文が引用している文献・引用されている文献へのリンクが表示される。また、論文自体は被引用文献の数によりランク付けされており、より多くの論文から引用されている論文はより高くランク付けされる。

- 類似論文の表示
  本システムでは、キーセンテンスなどから算出された類似度を元に、類似度の高い論文が表示される。これにより、直接的な引用関係は無いが内容的に類似している論文を参照することができる。

- 本文の表示
  CiteSeer では、ほとんどの論文を PDF や PS といったファイルの形で所有しているため、利用者は本文をそのまま手に入れることができる。

このように CiteSeer では、ただ論文を検索するだけでなく、その関連論文も検索可能であり、また本文を直接参照できるので、サーベイにかける時間を大幅に短縮することができる。

Google Scholar

Google Scholar は Google のデータから特に学術文献のみを抽出し、さらに論文に特化した検索機能を追加したシステムである。GoogleScholar の特徴を以下に紹介する。

- 大規模なデータベース
  公開されていないものの、Google のインデックスデータから取り出されていることから、非常に大規模であることが分かる。

- 引用・被引用文献の表示・
  検索した論文が引用している文献・引用されている文献へのリンクが表示される。また、論文自体は被引用文献の数によりランク付けされており、より多くの論文から引用されている論文はより高くランク付けされる。

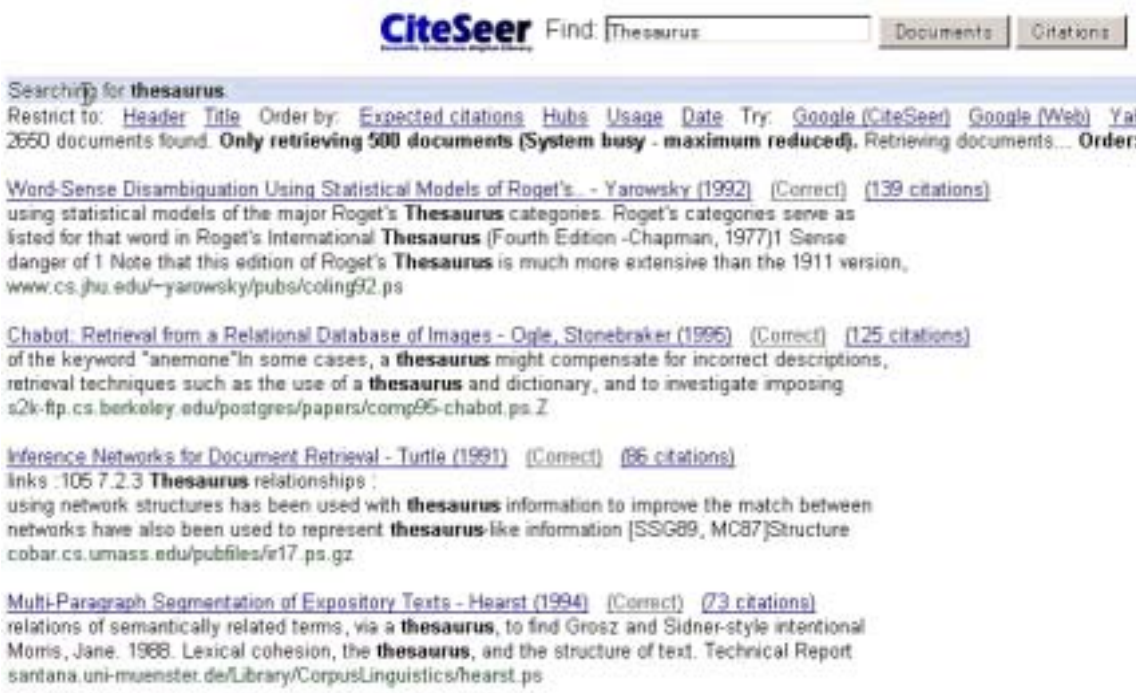

図 2.2: CiteSeer の画面

- 各論文検索サイトとの提携
  British Library Direct や ACM Portal と提携しており、それらのサイトの書誌情報を用いている。

- 図書館の検索
  各論文が収められている図書館を検索することができる。

このように Google Scholar は本家の Google と同じく非常に大規模なデータベースを持っており、また、各論文サイトとの提携や蔵書図書館の検索など、論文検索に特化した様々な機能が追加されているため、非常に使いやすいシステムとなっている。

PRESRI

PRESRI[58] は難波英嗣と東京工業大学奥村研究室のメンバが開発したシステムである。Web 上の日英論文データを収集して構築された。特徴は以下の通り。

- 中規模なデータベース
  18000 件の英語論文と 2000 件の日本語論文

- 引用・被引用文献とその引用タイプのグラフィカルな表示・
  論文の引用・被引用関係をグラフィカルに表示可能。また、本システムでは

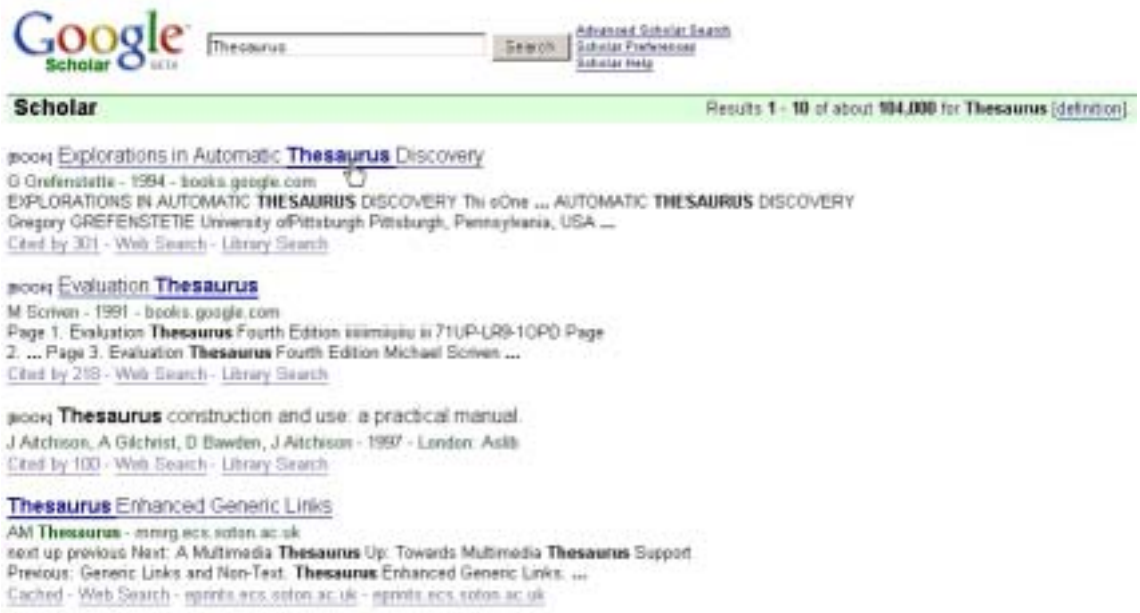

図 2.3: GoogleScholar の画面

引用の仕方を３種類に分類しており、論文がどのように引用されているかを知ることができる。

このように PRESRI は規模は大きくないものの、上記２つのシステムとは異なり引用関係とそのタイプをグラフィカルに表示するので、論文の関係が直感的に理解しやすくなっている。

これらのシステムは、Web 全体を対象としてデータのクローリングを行っているため、学会などが所有している論文検索システムと比べ、非常に広い範囲をカバーすることができる。また、内容の解析や論文間の関係性の解析なども行っており、あるキーワードに関連した論文やある論文に関連した論文など、より利用者の要求に沿った検索を行うことができる。例えば、GoogleScholar では、キーワードとの関連の強さと被引用数を元にした重要度により論文をランキングしている。CiteSeer でも、被引用数を元にした論文のランキングを行っており、また、ある論文と内容的に関連の強い論文を表示する機能もある。そのため、これらのシステムでは、あるキーワードから論文を検索する場合に、他のシステムを比べて読んでおくべき重要な論文を手に入れやすいと考えられる。PRESRI は引用文献の引用タイプの分類を行っており、どの引用文献を読めばよいかの判断の材料とすることできる。

その反面、Web 全体のクローリングには時間がかかるため、これらのシステムでは新しい論文に対応しにくい。また、被引用数を元にランキングするため、新しい論文が検索結果の上位になりづらい、という問題点がある。

2.2 従来システムの問題点

既存のサーベイ支援システムは、サーベイ作業をする上で有用なシステムであるが、完全な者ではなく、利用する上でいくつかの問題点が存在している。以下では、その問題点を指摘する。

2.2.1 検索キーワードの選択

GoogleScholar や CiteSeer などのシステムを利用すれば、あるキーワードに関連する論文を得ることができる。しかし、自分で選んだキーワードを用いて検索しても、目的としているテーマの論文が得られないことは多々ある。既存のシステムを利用して自分が望むような論文を検索結果として得るためには、適切な検索キーワードを探す必要がある。しかし、何の手がかりもなく適切な検索キーワードを探すというのは、非常に大変な作業である。より効率的に既存のシステムを利用するには、質問応答システムや検索クエリの拡張など、情報検索の技術を用いた支援を行う必要がある。

2.2.2 収集した論文の整理

我々が Web 上で検索する際は、キーワードを元に検索を行う場合と OpenDirectory[5]のようにあらかじめ分類されたディレクトリを元に検索を行う場合がある。また、サーベイ作業では、論文を収集した後、それらを有効に活用するために、論文集合を詳細な分野に分け、整理する必要がある。

このように、論文を整理し、分類する機能というのはサーベイ支援システムにおいて重要な機能である。しかし、既存のサーベイ支援システムでは、論文の収集を主な目的としており、収集した論文を整理する機能などは整備されていない。そのため、このような技術は今後必要になってくると考えられる。

2.3 提案システムの概要

本論文では、目的とする論文を検索するための検索キーワードの選定、及び収集した論文の整理・分類、という既存システムで解決されていない 2 つの問題点を解決するために、それぞれについて提案を行う。

[5]http://dmoz.org

2.3.1 検索精度向上のためのクエリ拡張

一般に情報検索の分野において、目的とする検索結果が得られるように検索キーワードを人手で決定することは難しい。しかし、得られた検索結果が目的に合致しているかどうかは、人間しか判断することができないため、検索キーワードを自動的に決定するのは困難である。

そこで、検索クエリ拡張の手法を用いて、検索を支援することを試みる。クエリ拡張とは、クエリの各単語や内容に関連する語を新たに検索クエリに付け加えることで、クエリとの単純マッチングでは検索できない文書を探し出す手法である。このクエリ拡張を用いて論文検索システムを利用することで、より効率的な論文検索を行うこと目指す。

2.3.2 時系列を考慮した論文クラスタリング

ある分野についての研究をきちんと把握するためには、その分野の論文を整理し、分野の変遷を把握することが必要である。しかし、自ら収集した論文を整理し、分野ごとに分類するのは非常に大変な作業である。

そこで、文書分類や文書クラスタリングの技術を用いてある論文集合をきちんと分野に分類することを目指す。ただし、研究分野は、それぞれの分野は成長、縮小するだけでなく、既存の分野が分裂したり、新たな研究分野が発生したりと時系列的に変化していく。しかし、既存の文書分類ではそのような時系列変化を考慮していないため適切な分類を行うことはできない。本論文では、そのような分野の発生や分裂を表現できるような論文分類技術を提案する。

第3章　Webを用いた検索クエリ拡張による関連論文検索

3.1　はじめに

クエリ拡張とは、元となる検索クエリに関連する語を検索クエリに追加することで、検索の精度をあげる技術である。クエリ拡張に関しては様々な研究が行われているが、いずれの手法も専門語のシソーラスや概念辞書などの専門分野に特化した知識ベースを用いて、元々のクエリに関連した語を取り出すことで、クエリの拡張を行っている。すなわちクエリ拡張では、関連語を取得するための知識ベースが不可欠であり、その精度は知識ベースに大きく影響される。ここでは、そのような知識ベースのうち、特にシソーラスに注目する。

代表的なシソーラスとしては、図3.1のようなWordNet[34]やEDR[60]など、長い年月を掛けて人手で構築したものがあげられる。しかし、こういったシソーラスは作成するのが手間がかかり、また日々現れる新しい語に対応するのが大変である。一方でシソーラスを自動的に構築する研究が以前から行われている。多様な語を扱うクエリ拡張では、シソーラスを自動で構築する、もしくは既存のシソーラスを自動で追加修正する手段が有効である。特に専門用語のシソーラスなどは、その分野の論文集合をコーパスとし、解析することで、シソーラスを自動構築することも多い。

シソーラスの自動構築は、シソーラスを構成する語の収集、語の関連度の算出と、その関連度を使った関連語の同定という段階に分けられる[33]。2語の関連度は、コーパス中の共起頻度を用いて求めることができる[5]。これまでの研究では、コーパスとして新聞記事や学術文書が用いられることが多かった。それに対し、近年ではWebをコーパスとして用いる手法が提案されている。Kilgarriffらは、Webをコーパスとして用いるための手法やそれに当たっての調査を詳細に行っている[22]。佐藤、宇津呂らはWebを用いた関連度の指標を提案している[59]。

Webには、新聞記事や論文といった従来からある整形された文書のみならず、日記や掲示板、ブログなど、よりユーザの日常生活に関連したテキストも数多く存在している。世界全体で80億ページを超えるWebは、間違いなく現時点で手に入る最大のコーパスであり、今後も増え続けるだろう。Kilgarriffらが議論しているように、Webの文書が代表性を持つのかといった議論はこれからも重要にな

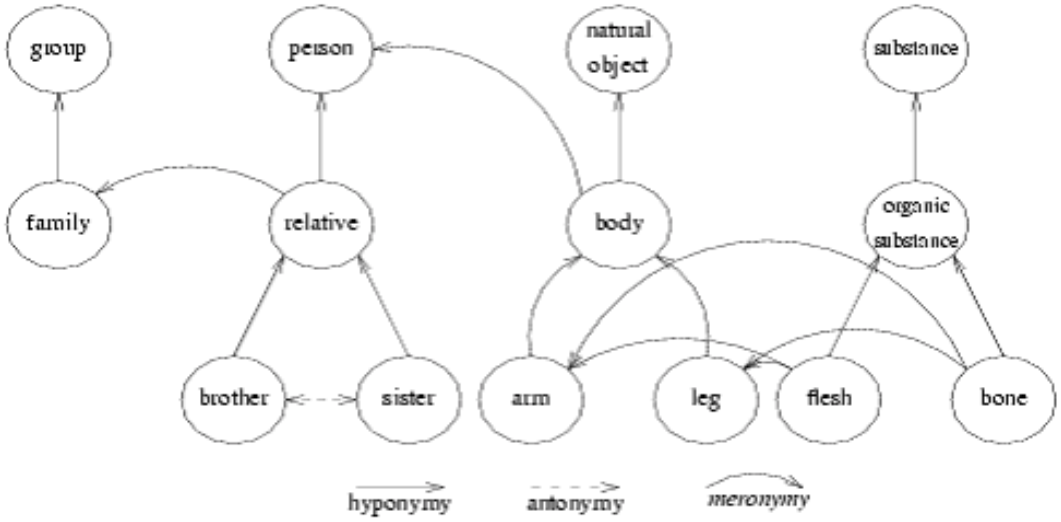

図 3.1: WordNet の構造

るが、Web はコーパスとしての大きな可能性を秘めていると著者らは考えている。Web をコーパスとして扱う際にひとつの重要な手段になるのが、検索エンジンである。これまでに多くの研究が検索エンジンを用いて、Web 上の文書を収集したり、Web における語の頻度情報を得ている [51, 17]。しかし検索エンジンを用いる手法とコーパスを直接解析する手法には違いがあるため、従来使われてきた計算指標がそのまま有効に働くとは限らない。

本論文では、Web を対象とし、検索エンジンを用いて検索クエリの関連語によるシソーラスを構築する手法を提案する。まず、検索エンジンを用いて検索クエリと関連の強い語 (以下、関連語と呼ぶ) を取得する。そして、それらの語を用いてクエリ拡張に利用可能なシソーラスを構築する。特に、検索エンジンを大量に使用すること、統計的な処理を行うことが特徴である。

また、シソーラスには WordNet を初めとする図のような階層構造をもつシソーラスと日本語大シソーラス [62] を初めとする図のような各語の関連語をフラットに並べたシソーラスが存在している。この２つを比較すると、階層構造を持つシソーラスの方が適用範囲が広い反面、自動構築のコストが高く、また階層構造の精度が低いという問題点がある。本研究では、検索クエリ拡張のためのシソーラスであり、階層構造を必要としていないので、関連語を並べたシソーラスを構築の対象とする。

3.2　検索エンジンを用いた検索クエリの関連語の取得

本節では、シソーラスの構成語として、検索クエリに用いる語に強く関連している語を取得する手法について説明する。本手法は、図 3.2 のようにもとの検索クエリの語から関連語候補を取得する段階と、得られた関連語候補を順位付けを行う段階の２つの段階の分けられる。いずれも検索エンジンを利用しているのが大きな特徴である。なお、この手法は佐藤ら [59] が用いている手法と同じである。

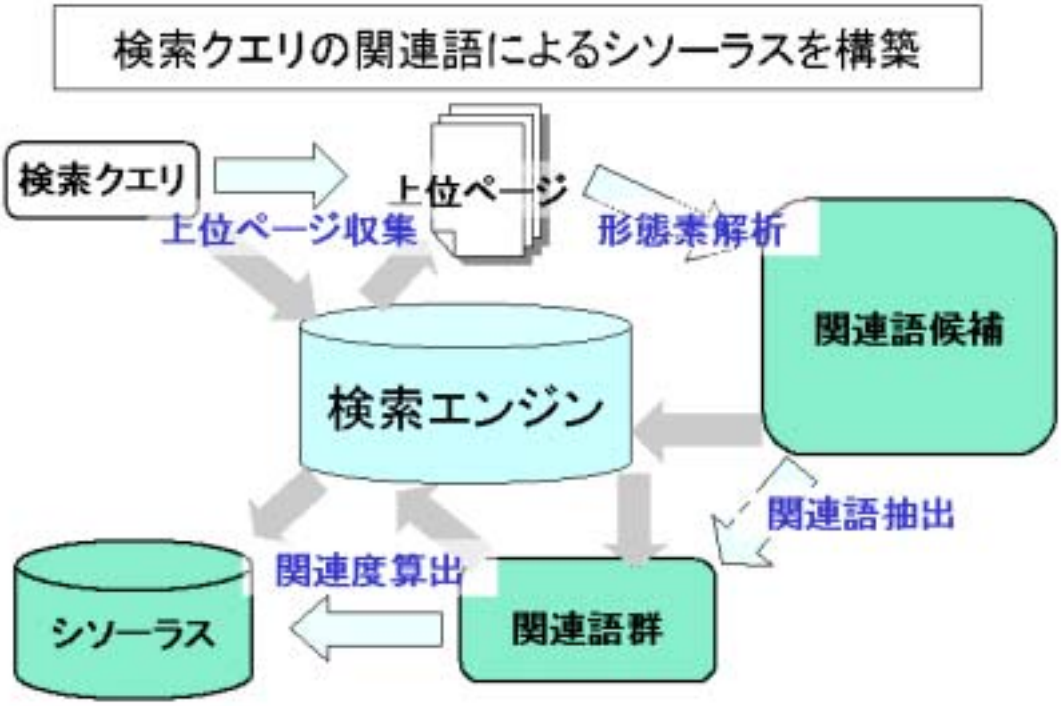

図 3.2: Web を用いたシソーラスの構築手法

3.2.1　関連語候補の取得

提案手法では、検索エンジンの上位ページに出現する語を検索クエリの連想語候補を取得する。

検索エンジンは、Web 上ページを検索クエリーとの関連の深さによってランクづけし、その上位 n ページを検索結果として返すことを目的としている。以前は、ランクづけの精度は低く、望むような結果が得られなかった。しかし近年、Google の PageRank[38] を初めとする検索エンジンの技術の向上により、かなりの精度で

検索クエリーと関連の強いページが上位にランクされるようになっている。明確に検索結果の効果を示す実験などは行われていないが、Google や Yahoo などの人気のある検索エンジンは、その利用頻度の高さから考えて、かなりの精度で検索クエリーと関連のあるページが上位にランクされていると考えられる。また佐藤らにより、検索エンジンの結果をコーパスとすることで、専門用語の関連語が収集できることが示されている [59]。つまり、検索結果の上位にランクされるページは、検索クエリーと強い関連を持ったトピックを扱っているとみなすことができる。

あるトピックを扱うページには、そのトピックに関連している語が複数出現していると考えられる。そこで次のような仮定をおくことができる。

「検索結果の上位にランクされるページには、
検索クエリーとの関連が強い語が出現している」

この仮定に基づき、検索クエリを検索クエリーとして検索を行い、その検索結果の上位ページ群に一定以上の頻度で出現する語を検索クエリの連想語候補として取得する。

3.2.2 関連語の重み付け

検索結果の上位ページから連想語候補を獲得した。しかし、ある語が「検索エンジンの上位ページに出現する」ということは、「検索クエリと関連している」ことの十分条件ではない。また、連想語候補群と検索クエリの関連度がすべて同じではない。そこで、各連想語候補と検索クエリの関連度を測り、その関連度の強い語を検索クエリの連想語とみなすこととする。

2 語の関連度を測る手法は、従来は新聞記事コーパスや論文コーパス中における 2 語の共起情報を用いる手法が一般的であった。一方、近年は検索エンジンを通じて Web 全体をコーパスとして用いる手法が増えてきている。従来手法では扱える語が限られるのに対し、検索エンジンを用いた手法では Web 上に出現する語ほとんど語を扱うことができる。検索エンジンの上位ページから得られる連想語候補群には、非常に広範囲の語が含まれる可能性が高いため、本手法では検索エンジンを用いて 2 語の関連度を測ることとする。

検索エンジンを用いて 2 語の関連度を測る場合は、2 語の単独ヒット件数および共起ヒット件数を利用する。検索エンジンでは、「”語 w_i”」をクエリーとして検索すると、語 w_i のヒット件数が得られる[1]。検索エンジンは非常に多くのページをクローリングしているため、このヒット件数を語 w_i の Web 全体での出現頻度と近似できる。同様に「”語 w_i 語 w_j”」をクエリとして検索れば、語 w_i と語 w_j の Web 上での共起頻度が得られる。これらの値に表 3.8 の確率手法の計算指標を適

[1]ダブルクォーテーションで囲んでいるのは、2 語以上からなるフレーズに対しても適切に処理するためである。

用することで、2 語の関連度を計算することができる。例えば、相互情報量は、語 w_a の出現確率を $p(w_a)$、語 w_b の出現確率を $p(w_b)$、語 w_a と語 w_b の同時出現確率を $p(w_a \cap w_b)$ とすると、

$$\begin{aligned} MI(w_a, w_b) &= \log \frac{p(w_a \cap w_b)}{p(w_a)p(w_b)} \\ &= \log \frac{Nn(w_a, w_b)}{n(w_a)n(w_b)} \end{aligned} \tag{3.1}$$

と表される。ここで $n(w_a)$ は語 w_a をクエリーとしたときのヒット数、$n(w_a, w_b)$ は「語 w_a 語 w_b」をクエリーとしたときのヒット数であり、また、N は検索エンジンのクロールした全ページ数である。Baroni らは N を 3 億 5 千万ページとしているが、2005 年末現在では、Google は約 100 億ページ、AltaVista は約 120 億のページである。ここでは $N = 100 \times 10^8$ とした。

従来の検索エンジンを用いて 2 語の関連度測る手法では、条件付き確率や相互情報量、Jaccard 係数などが計算指標として用いられている。しかし、条件付き確率は相互情報量、Jaccard 係数などより値が精度が低いことが知られている。そこで、相互情報量と Jaccard 係数を予備実験により比較し、より精度の高い計算指標を適用する。

3.3 シソーラスの構築

本節では、Web 上の情報を用いて語の関連度を測る手法を提案する。

3.3.1 検索エンジンのヒット件数の利用と従来手法の問題点

3.2 節では、検索クエリと関連語候補との関連度を相互情報量と Jaccard 係数を用いた。しかし、これらの計算指標は、検索エンジンのヒット件数を用いる際には問題がある。

具体的な例を使って説明しよう。関連度を計りたい語を、例えば「インク」「インターレーザー」「プリンタ」「印刷」「液晶」「Aquos」「TV」「Sharp」の 8 語とする。これらの語群は、Epson のプリンタであるインターレーザーに関する語と、Sharp の液晶 TV である Aquos に関する語であり、各語の関連度を得ることで、2 つのグループを適切に分けたいと仮定する。

表 3.1 は、語群の各語の単独ヒット件数、表 3.2 には、語群中の 2 語を共起ヒット件数を行列形式にしたもの、表 3.3 は相互情報量を示す。。

相互情報量を例に取ると、この計算指標は「出現確率に影響を受ける」という特徴を持つ。この特徴は式 (3.1) を次式のように書き換えるとわかりやすい。

$$MI(w_a, w_b) = \log p(w_a|w_b) - \log p(w_a) \tag{3.2}$$

$p(w_a|w_b)$ は語 w_b が出現するときに語 w_a と語 w_b が共起する条件付き確率を表す。$p(w_a|w_b)$ が等しい場合は、$p(w_a)$ の出現確率が小さいほど相互情報量は大きい値になる。この特徴自体は「共起する確率が同じなら、出現確率の低い語と共起する方が関連性が強い」と考えられるので、問題がない。しかし、検索エンジンにおいては語によって出現頻度に大きなばらつきがあり、また全事象を表す N が非常に大きいために出現確率の違いによる影響が大きくなり過ぎてしまう。実際に表 3.1 の語のヒット件数と表 3.3 の各行との相関係数は -0.65 となり、相互情報量と語の出現確率に強い負の相関があることが分かる。それに対し、表 3.2 の共起ヒット件数と表 3.3 の相互情報量のとの相関係数は -0.15 となり、あまり相関がないことが分かる。

このように、従来用いられてきた相互情報量は語の出現確率に影響を受けるため、関連度を測る際に各語の出現確率に数千倍、数万倍といった開きがある場合、値の信頼性は低くなるという問題がある. これは、Jaccard 係数や dice 係数など他の類似度の指標についても当てはまる。

表 3.1: 語単独でのヒット件数

プリンタ	印刷	インターレーザー	インク	液晶	TV	Aquos	Sharp
17000000	103000000	215	18900000	69100000	1760000000	2510000	186000000

表 3.2: ２語でのヒット件数の行列

語/語	プリンタ	印刷	インターレーザー	インク	液晶	TV	Aquos	Sharp	合計
プリンタ	0	4780000	273	4720000	4820000	4530000	201000	990000	20041273
印刷	4780000	0	183	4800000	6520000	8390000	86400	1390000	25966583
インターレーザー	179	183	0	116	176	65	0	0	813
インク	4720000	4800000	116	0	3230000	10600000	144000	656000	24150116
液晶	4820000	6520000	176	3230000	0	13900000	903000	4880000	34253176
TV	4530000	8390000	65	10600000	13900000	0	1660000	42300000	81380065
Aquos	201000	86400	0	144000	903000	1660000	0	1790000	4784400
Sharp	990000	1390000	0	656000	4880000	42300000	1790000	0	52006000
合計	20041273	25966583	813	24150116	34253176	81380065	4784400	52006000	242582426

3.3.2 χ^2値を用いた関連度の指標の提案

本論文では、χ^2 値を使った関連度の指標を用いる。χ^2 値は、あるデータ集合内での統計的な偏りを表す指標であり、機械翻訳やコロケーション処理など、多くの手法で用いられている。語の関連度としては Curran らが用いている [8]。

表 3.3: 相互情報量行列

語/語	プリンタ	印刷	インターレーザー	インク	液晶	TV	Aquos	Sharp
プリンタ	0	4.195	7.504	5.878	4.602	1.303	4.740	2.029
印刷	4.195	0	5.302	4.093	3.103	0.117	2.094	0.567
インターレーザー	7.504	5.302	0	6.542	5.663	1.429	0.000	0.000
インク	5.878	4.093	6.542	0	4.096	2.047	4.301	1.512
液晶	4.602	3.103	5.663	4.096	0	1.021	4.840	2.222
TV	1.303	0.117	1.429	2.047	1.021	0	2.212	1.144
Aquos	4.740	2.094	0.000	4.301	4.840	2.212	0	4.534
Sharp	2.029	0.567	0.000	1.512	2.222	1.144	4.534	0

χ^2 値を関連度に用いるのは、語の出現頻度のばらつきによる影響を排除するためである。相互情報量や Jaccard 係数を関連度に用いる場合の問題点は、語の出現確率に大きな影響を受ける点である。この問題の解決策として、出現確率を適切に正規化するというアプローチが考えられる。χ^2 値では、語群を構成する語の出現頻度を正規化要素とし、値の正規化を行ったうえで、共起の偏りを算出するので、出現確率のばらつきによる影響を抑えることができる [56]。このため、値のばらつきが大きい検索エンジンのヒット件数を用いて関連度を算出する場合、χ^2 値を計算指標として用いることが適切であると考えられる。

対象とする語群の中で、共起の偏りを統計的に調べるために、1 つ 1 つの語について、語群内の他の語との共起頻度を標本値とし、「$w_i, w_j \in G$ が共起する確率は、語 w_i と語群 G 内の語が共起する確率と等しい」という帰無仮説をおいて検定を行う。語 w_i と語 w_j の実際の共起頻度を $n(w_i, w_j)$、語 w_i と語群 G の語との共起頻度の和を $S_{w_i} = \sum_k n(w_i, w_k)$、全ての共起頻度の和を $S_G = \sum_{w_i \in G} S_{w_i}$ とするとき、語 w_i と語 w_j に関する χ^2 値は次式で表される。

$$\begin{aligned} \chi^2(w_i, w_j) &= \frac{n(w_i, w_j) - E(w_i, w_j)}{E(w_i, w_j)} \\ E(w_i, w_j) &= S_{w_i} \times \frac{S_{w_j}}{S_G} \end{aligned} \tag{3.3}$$

$E(w_i, w_j)$ は語 w_i,w_j の共起頻度の期待値を表している。例えば、語 w_i を「プリンタ」、語 w_j を「インターレーザー」とすると、$n(w_i, w_j)$ は 179、$S_{w_i} = 20041273$、$S_{w_j}/S_G = 813/242582426$ となる。表 3.4 は、表 3.2 から計算された χ^2 値行列である。表 3.4 では、「プリンタ」は「印刷」や「インク」と偏って共起している。また、「インターレーザー」は「プリンタ」との共起が、「Aquos」は「Sharp」との共起が強いなど、良好な結果となっている。

表 3.4: χ^2 行列

語/語	プリンタ	印刷	インターレーザー	インク	液晶	TV	Aquos	Sharp
プリンタ	0.000	3235887	630.8	3721225	1399572	0.000	0.000	0.000
印刷	3235887	0.000	105.8	1897753	2220688	0.000	0.000	0.000
インターレーザー	630.8	105.8	0.000	15.19	32.63	0.000	0.000	0.000
インク	3721225	1897753	15.19	0.000	0.000	770371	0.000	0.000
液晶	1399572	2220688	32.63	0.000	0.000	505007	76566	0.000
TV	0.000	0.000	0.000	770371	505007	0.000	1882	35404428
Aquos	0.000	0.000	0.000	0.000	76566	1882	0.000	569512
Sharp	0.000	0.000	0.000	0.000	0.000	35404428	569512	0.000

3.4 自動構築したシソーラスを用いたクエリ拡張

検索エンジンによる検索クエリからの関連語の取得および χ^2 値による関連度の算出を用いたシソーラスの構築、さらにそのシソーラスを用いたクエリ拡張の手順を以下に示す。

1. 基となる検索クエリ (以下、検索クエリと呼ぶ) をクエリーとして検索エンジンを利用し、その検索結果上位 n ページを取得する。

2. 得られた上位 n ページを形態素解析し、5-gram までを単語として取り出す。ただし、1 つのページにしか出現しない語はほとんど関係ないと考えられるので切り捨てる。

3. 2. で得られた語群に対して、各語の単独ヒット件数と各語と検索クエリとの共起ヒット件数を、検索エンジンより取得する。

4. 3. で得た頻度情報を基に、各語と検索クエリとの相互情報量を算出し、その上位 299 語を取り出す。その 299 語と検索クエリを合わせて語群 G とする。

5. 式 3.3 により、語群 G 内での各語同士の χ^2 値を算出し、それを 2 語間の関連度とする。

6. (5) で算出した関連度をもとに、もととなった検索クエリと関連の強い語を選び出し、それを関連語とする。

7. 取得した関連語群の 1 語と検索クエリを組み合わせた計 2 語を拡張クエリとする。つまり、取得した関連語の数だけ拡張クエリができる。この拡張クエリを用いて検索を行う。

3.5 具体例と評価

提案手法では、検索クエリからの関連語の取得、検索エンジンを用いた関連度指標、の 2 つにより、新たな検索クエリ拡張の手法を提案している。そのため、提

案手法を評価するためには、2つの手法の評価、およびその2つを合わせた検索クエリ拡張の評価を行う必要がある。しかし、検索クエリと関連している関連語というものは実際は無数にあるため、適切な正解データを得ることができず、関連語の取得を評価することは難しい。そこで本論文では、関連度の指標の適切の評価および全体を用いた検索クエリ拡張の効果を確かめることで、提案手法全体の評価を行いたい。

3.5.1 関連度指標の評価実験手法

語の関連度を評価する手法として、WordNet や EDR など人手で構築された既存のシソーラスと比較する方法 [19, 7]、綿密に作られたアンケートや語の分類タスクを人が行い、その結果と比較することでシソーラスの適切さを評価する方法 [44, 18] がある。しかし、後者は非常にコストが高く、またある程度の量のデータを集める必要があるため実現は難しい。そこで本評価実験では、図 3.3 のように既存のシソーラスである WordNet と出力結果を比較することで、関連度の評価を行いたい。

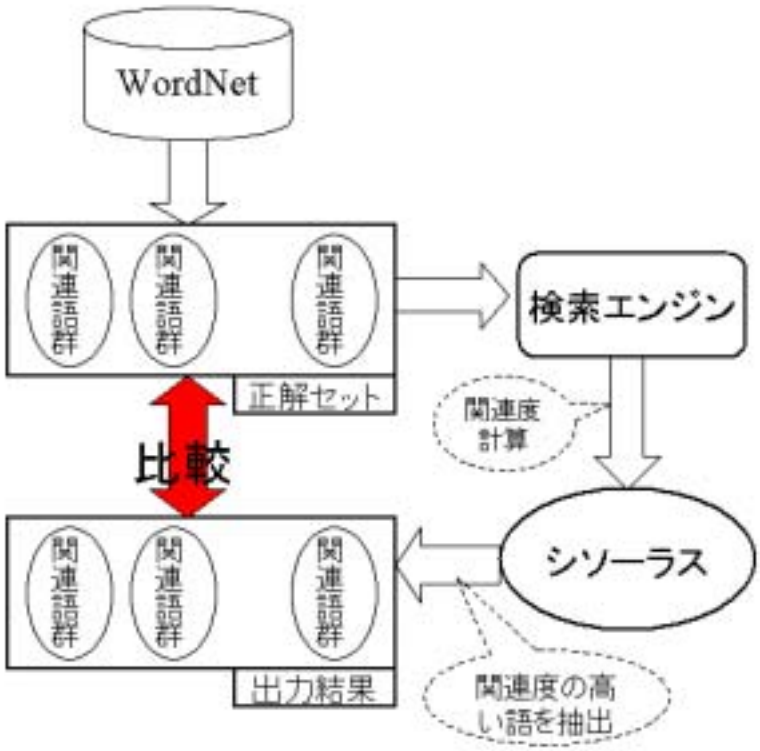

図 3.3: WordNet との比較による関連度評価実験

まず、WordNet から正解データを作成する。WordNet は階層構造を持っており、それにより上位語、下位語などが取得できる。その中でも同じ上位語を持つ同位語同士は関連語として扱えると考えられる。例えば、「バイオリン」や「ビオラ」は「弦楽器」という上位語を通して関連している。そこで今回は同じ上位語を持つ同意語同士を関連語を正解データとして取り出した。正解データを作成する手順を以下に示す。

1. 一つの意味で下位語を 10 語以上持つ語を選ぶ。ただし、あまり WordNet の上位層にいる語は抽象的すぎるため選ばない。(例：”もの”、”事物”)

2. その語の下位語の中から 10 語を選びを 1 つの関連語群 W_i とする

3. 1.,2. を $i = 1.....10$ までについて繰り返す

以上のような手法を用いて、「コンピュータ」「学問」「宝石」「職業」「芸術」「内臓」「文芸」「食品」「飲み物」「スポーツ」の下位語を関連語として取り出した。なお、これらの語はそれぞれ、「物質」「活動」「位置」の下位に属する語を用いている。

図 3.4 に抽出した関連語の例を示す。

このように算出した正解データ $W_i(i = 1...10)$ を用いて評価実験を行う。基本的な考え方としては、まず元々の正解データを語群として与える。そして、その語群の関連度を計算し、正解データにおいて関連語となっている語同士の関連度が高くなっているかによって評価を行う。ただし、用いた正解セットの場合に偶然精度が高いことも考えられるので、10 分割交差検定を行う。評価実験の手法を以下に示す。

1. 正解データから W_i から 9 つの関連語群を選ぶ。

2. 9 つの関連語群に含まれる全ての語について共起ヒット件数を取得する

3. χ^2 値を算出し、関連度とする。

4. 各語 w ごとに関連語が高い語を 9 つ選び、それを語 w の関連語と比較し、適合率・再現率を計算する。

5. 1.〜4. を繰り返す。

シソーラスの評価は、適合率と再現率の簡単な算出例を表 3.5 に示す。この場合、手法 1 による出力は 3 語中 2 語が正解であるので適合率は 0.667、正解データ 4 語のうち 2 語が手法 1 により出力に含まれているので、再現率は 0.5 となる。

提案手法では、χ^2 値を用いて関連度を算出しているが、比較手法としては、Jaccard 係数と相互情報量を用いた。また、関連度を測るための検索エンジンとしては Google[2]を用いた。

実験結果を表 3.6 に示す。

[2]http://www.google.com

図 3.4: WordNet から抽出した関連語群

表 3.6 を見ると、いずれのセットにおいても 3 つの手法の中では χ^2値 が優れている。その次は相互情報量が優れており、Jaccard 係数の値は最も低くなっている。

これより、検索エンジンのヒット件数を用いた語の関連度の算出には、計算指標として χ^2値 が優れていることが分かる。また、χ^2値 が使えないような 1 対多の関連度を測る際は、相互情報量を用いることが適切であるといえるだろう。

このようにして χ^2 値を用いることで、6 割程度の精度で WordNet を再現することができる。ただし、提案手法は全ての語・全ての語の関係に均等に作用するわけではなく、提案手法で得意とする範囲と苦手とする範囲がある。実験結果を細かく分析すると、関連語の同定に成功する場合、失敗する場合は図 3.5 のようになる。

成功例をみると、一義的な語や具体語、特に「宝石」に関連するのカタカナ語は 100%の精度で抽出することができている。逆に失敗する場合を見ると、まず「ホームページ」や「掲示板」といった Web ページ内を表現する語の抽出に失敗している。また、「装飾」といった多義的で抽象的な語の抽出に失敗している。

つまり、文脈に関係なく出現する Web 特徴語や多義語、抽象語などの語は提案手法で扱うことが難しくなっている。

表 3.5: 評価実験の例 (ヴァイオリン)

	関連語	適合率	再現率
正解データ	ビオラ, チェロ, 笛, ギター		
手法 1	ビオラ, チェロ, ビール	0.67	0.5
手法 2	ビオラ, チェロ, ピック, 笛, ギター	0.8	1.0

表 3.6: WordNet との比較実験結果

セット	χ^2 値	相互情報量	Jaccard 係数
セット 1	0.6039	0.5825	0.49625
セット 2	0.65513	0.58375	0.49125
セット 3	0.57564	0.54875	0.48
セット 4	0.64875	0.59375	0.50375
セット 5	0.58077	0.5525	0.47375
セット 6	0.59744	0.5575	0.47875
セット 7	0.6	0.57	0.495
セット 8	0.4975	0.4725	0.415
セット 9	0.52821	0.52	0.44125
セット 10	0.5641	0.51625	0.4725
平均	0.585144	0.54975	0.47475

3.5.2 検索クエリ拡張の評価

本章で提案した検索クエリ拡張の手法を用いることで、実際に検索の精度が向上するか、という評価を行う。

本評価実験では、検索精度の評価セットとして TREC7 で用いられた OHSUMED によるデータセットを用いる。TREC とは Text REtrieval Conference の略であり、用意された検索対象のデータを用いてお互いの検索システムの精度を競い合う情報検索のコンテストである。アメリカ標準技術局 (NIST) により、1992 年から毎年開かれている。

OHSUMED はオレゴン州立健康科学大学が医学文献を集めたサイト MEDLINE から医学文献を収集したものであり、多くの情報検索の研究の評価に用いられている。大量の文書と 106 の検索クエリが準備されており、各検索クエリとそのクエリの検索結果として適切である文書が対応付けされている。今回の評価実験では、その中の 50 クエリを用いて評価実験を行う。

検索エンジンとしては、Salton らによる SMART[42] を用いる。SMART は検索タスクの評価実験によく用いられる検索システムである。

図 3.5: 関連語抽出の正解例・失敗例

データセット OHSUMED を用いて検索エンジン SMART を評価する手順を以下に示す。

1. 与えられた検索クエリ q_i を用いて検索を行う。
2. 検索結果と元々q_i に関連づけられていた論文とを比較
3. 適合率・再現率を算出する。

これはあくまで検索エンジン SMART の評価手法であるが、これを比較手法として用いる。拡張クエリを用いて SMART で検索を行うことで、この比較手法の精度よりも高い精度を得ることができれば、拡張クエリの有効性を示すことが出来る。拡張クエリを用いた検索手法を以下に示す。

1. 検索クエリ q_i をもとに Web を利用して、拡張クエリ E_{ij} を取得する。
2. 元のクエリ q_i と拡張して得られた拡張クエリ E_ij を用いて検索を行う。この際、拡張クエリは $E_ij(j = 1....20)$ と 20 個あるので、それぞれ一つずつ、計 20 回検索を行う。

3. 20 回の検索の結果、得られた正解文書数を元に、精度、再現率を算出する。

比較手法は、拡張せずに与えられた検索クエリ q_i をそのまま用いる検索手法とする。まず、精度と再現率の相関グラフを図 3.6 に示す

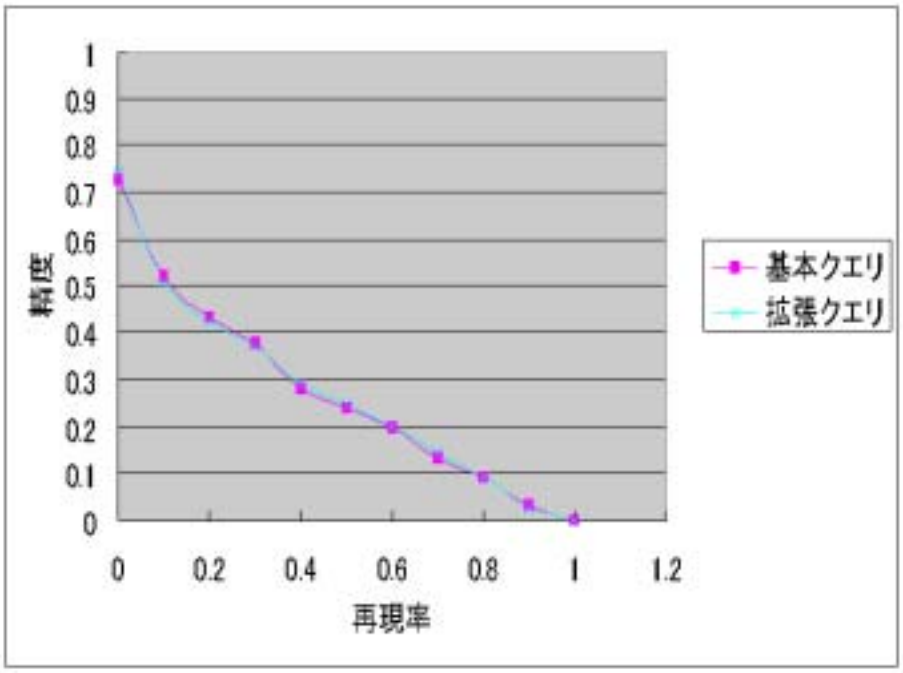

図 3.6: 検索クエリ拡張の評価実験結果

図 3.6 においては、再現率と精度がトレードオフの関係になっている。これは、検索結果の上位何件までを判定するかの問題であり、再現率をあげるためには、出きる限り多くの論文を検索結果から取り出す必要があるが、取り出す論文が多いほど、精度が下がってしまうためこのようなグラフになる。

図 3.6 において、拡張テストクエリとテストクエリの結果を比較すると、精度・再現率ともにほとんど変わっていない。これは、拡張テストクエリによる影響がないわけではない。出力結果の内訳を見ると、拡張クエリを用いることで精度が向上するクエリと、低下するクエリがあるため、トータルでうち消し合って、結果的に 2 つの手法はほぼ同じ性能となっている。

次に表 3.7 に、性能が向上する場合と低下する場合に分けた場合のそれぞれの精度・再現率の平均値を示す。ここれは、性能が変わらない場合は除いている。

表 3.7 をみると、性能が向上する場合はテストクエリの精度平均が 0.06、拡張クエリの精度平均が 0.22 であり、テストクエリでは低かった精度が大幅に向上している。逆に、性能が低下する場合に注目すると、テストクエリの精度平均が 0.37、拡張クエリの精度平均が 0.37 であり、元々テストクエリである程度の高かった精度が少しだけ低下している。

表 3.7: WordNet との比較実験結果

	精度向上	精度低下
基本クエリ	0.067	0.37
拡張クエリ	0.20	0.30

これより、テストクエリですでに適切な結果が得られている場合に拡張クエリを用いると、適切な結果が返されない場合が多いことがわかる。逆にテストクエリでの適切な結果が得られていない場合に拡張クエリを用いると、適切な結果が得られる可能性があがることがわかる。

以上の結果から、まずテストクエリを用いて検索によって適切な検索結果が得られなかった場合に、拡張クエリを用いることが有効であることがわかる。

3.6 議論

本章では、シソーラスを用いた検索クエリ拡張により、Web 上からの論文検索の精度の向上を目指した。

Web 上の情報を用いる際、・上位ページを用いた検索クエリの拡張、・検索エンジンのヒット件数を用いた関連度の算出と 2 通りの方法で検索エンジンを利用している。そのうち、ヒット件数を用いた関連度が有効であることは評価時実験で示したが、上位ページを用いた検索クエリ拡張の評価については行っていない。しかし、検索クエリ拡張自体の精度が向上していることから、上位ページからの語の収集することで、検索クエリと関連のある語が収集できていると推測される。また、現在までに Web 上の検索エンジンの利用率の高さからも、検索クエリに何らかの形で関連したページが検索結果として得られていると考えられるため、上位ページからの語の収集は有効であると推測できる。

また、本実験で Web を用いた関連度では、抽象的な概念の語や Web に特徴的に出現する語を対象とすることを苦手としていることが明らかになった。しかし、本研究で対象としているのは論文であり、専門用語を扱うことが多いため、この点は問題が無いと言える。

今回の実験で、元々のクエリで適切な結果が得られない場合に構築したシソーラスを用いたクエリ拡張が有効であることが分かった。つまり、実際に運用する場合は、検索結果を見てユーザーがクエリ拡張を行うかを判断する必要がある。これは、自動化という観点からみれば欠点あるが、現在稼働しているクエリ拡張の多くが対話型クエリ拡張であるため、特に問題がないと考えられる。

また、ローカルでの検索と Web の検索はデータ量、検索システムなどの違いがあるため、直接比較することは難しい。しかし、ローカルでの情報検索においては確かに精度が向上しており、本手法の有効性を示す参考データとしては十分で

あると筆者らは考えている。

今回は、Web 上の情報を利用して、論文検索の精度を向上させる手法について提案を行った。今後は、提案手法の関連度をなどを用いて、関連文献の検索を支援するような手法を提案していきたい。

3.7 関連研究

3.7.1 検索クエリ拡張の従来研究

検索クエリ拡張とは、ユーザーによる検索クエリに対して適切な検索クエリを与えることで、情報検索の精度と再現率の向上を目指す手法である。これまで様々な研究が行われているが、大きく分けて 2 つのアプローチがある。一つは、ユーザーから得られるフィートバックを利用するアプローチである。まず、ユーザーから得られる検索結果の関連性判断や検索履歴などを用いて、クエリとそのクエリの検索結果に対する文書の適切性の判定情報を得る。そしてその情報を元により適切なクエリを選択する方法である。これは Salton ら [42, 43] や [16] などで提案されている手法である。確かにこのような一般ユーザーからのフィードバックを用いることで、検索性能を向上させるようなクエリを選択することができる。

もう一つは、シソーラスや概念辞書などの知識ベースをもとにクエリの関連語を取り出すことで、検索クエリの拡張を行う手法である。検索クエリ拡張のための知識ベースを構築する手法としては、特定分野の文書を訓練データとして利用することでドメイン特化型のクエリ拡張を行う手法 [12, 37] がある。また、手動で用意された適合文書や適合文書であると見なされた文書を利用する手法があり、その中でも特に Robertson の方法 [41] は、多くの研究で用いられている [57, 25]。また、検索クエリ専用の知識ベースを用いるのではなく、既存のシソーラスや、自ら構築したシソーラスを用いることで検索クエリ拡張を行っている手法もある [45, 15]。これらのシソーラスの構築手法については、次節で解説する。

これら 2 つを比較すると、通常の検索システムにおいて、ユーザーから十分なフィードバックを得ることは難しい [49] そのため、本研究ではシソーラスを構築するアプローチを用いている。しかし、シソーラスも知識ベースを共に訓練データとして利用するテキスト群により扱える範囲が限られてしまう。特に様々な分野での専門的な知識を必要とする論文検索の検索クエリ拡張においては、訓練データが大きな問題となる。本研究では、Web シソーラスを持っているため、これらの問題を解決することができる。つまり、近年の Web の発達により非常に多くの分野の論文が Web 上の存在している。そのため、Web から知識を獲得することで、様々な分野の知識をカバーしているのである。これにより、様々な学問分野に対応した検索クエリ拡張が可能となるのである。

また、検索クエリ拡張は拡張クエリの追加方法で分類すると、対話型検索クエリ拡張と自動検索クエリ拡張に分けられる [41]。対話型検索クエリ拡張とはクエリ

拡張の候補語をユーザーに提示し、その中からユーザーが目的とする語を選ぶ手法である。つまり、自動化されているのはクエリ拡張の候補語の提示までである。それに対し、自動検索クエリ拡張とは、クエリ拡張に用いる語の決定までが自動された手法である。Koenenmann により、対話型クエリ拡張の方が自動検索クエリ拡張よりも精度が高いことが示されている [23]。本研究では、評価のために自動検索クエリ拡張を行ったが、実際に用いる場合は、対話型検索クエリで用いる方が優れていると考えられる。

3.7.2 2語の関連度を測る従来研究

語の関連性を自動的に得る方法は、これまでにさまざまな研究が行われている。コーパス中での語の共起情報をもとに語の関連度を測る指標として、様々なものが提案され用いられており [5, 52, 44, 8]、それらは大きく 2 つに分けられる。1 つは単語ベクトルを用いたベクトル空間手法である。これは、単語を多次元ベクトル空間の単語ベクトルで表現し、それぞれの単語ベクトルを比較することで関連度を測る手法である。ベクトル空間手法では、表 3.8 のようにベクトルの内積をもとにした計算指標が用いられている。表 3.8 において、x_i, y_i はそれぞれ単語ベクトル $\vec{x}, \vec{y}$ の i 番目の要素を表す。なお、overlap 係数はバイナリベクトルにしか用いることはできない。単語ベクトルの要素の取り方は研究によって様々であり、各文書への出現頻度を要素とするベクトルや各単語との共起頻度を要素とするベクトルなどが考えられる。ただし、独立な事象の確率は足し合わせることができないため、内積を用いる関連度では、語の出現確率を単語ベクトルの要素とすることは不適切と考えられる。

もう 1 つはコーパス中での確率を用いる確率手法である。この手法では、2 語がコーパス中で共起する確率をもとに関連度を算出している。確率手法で用いられている計算指標を表 3.8 に示す。表 3.8 において、$p(w \cap w')$ は語 w, w' の共起確率を表し、$p(w \cup w')$ は語 w, w' のどちらかが出現する確率を表す。また f は [32] で定義されている関数であり、$f(w, r, w')$ は語 w, w' が r の関係を持って出現する頻度を、$f(*, r, w')$ は語 w' がいずれかの語と r の関係を持って出現する頻度を表す。これらの計算指標は、ベクトル空間手法で用いられている指標を書き換えたものが多い。また、単語同士の共起確率ではなく、各単語が他の語と共起する確率の確率分布関数の類似性を用いて関連度を算出する研究も数多く行われている [4, 2, 47]。確率分布関数を用いた類似度は、確率分布類似度（Distributional Similarity) と呼ばれる。類似した名詞は共通した動詞と共起すると仮定し、動詞との共起分布の類似性から関連度を算出している。

語の関連度が得られれば、関連度に基づいて語をクラスタリングすることで関連語が得られる。実際には、同じクラスタに分類された語同士を関連語や同義語

[2] [32] で提案されている手法
[3] [32] で提案されている手法

表 3.8: 類似度の計算指標

ベクトル空間手法		確率手法	
cosine	$\frac{\vec{x}\cdot\vec{y}}{\sqrt{\lvert\vec{x}\rvert\lvert\vec{y}\rvert}}$	相互情報量	$\log\left(\frac{p(w\cap w')}{p(w)p(w')}\right)$
dice	$\frac{2(\vec{x}\cdot\vec{y})}{\sum(x_i+y_i)}$	dice	$\frac{2p(w\cap w')}{p(w\cup w')}$
Jaccard	$\frac{\vec{x}\cdot\vec{y}}{\sum(x_i+y_i)}$	Jaccard	$\frac{p(w\cap w')}{p(w\cup w')}$
overlap	$\frac{\lvert\vec{x}\cap\vec{y}\rvert}{min(\lvert\vec{x}\rvert,\lvert\vec{y}\rvert)}$	T 検定	$\frac{p(w\cap w')-p(w')p(w)}{\sqrt{p(w')p(w)}}$
Lin[3]	$\frac{\sum(x_i+y_i)}{\lvert\vec{x}\rvert+\lvert\vec{y}\rvert}$	Lin98A[4]	$\log\left(\frac{f(w,r,w')f(*,r,*)}{f(*,r,w')f(w,r,*)}\right)$

であるとしている。語のクラスタリングには分布クラスタリング (Distributional Clustering) が用いられることが多い。分布クラスタリングとは、類似した名詞は共通した動詞と共起すると仮定し、各語の動詞との確率分布の類似度に基づいて、データを結合もしくは分割していくクラスタリング手法である [40, 30, 9]。

これらコーパスから関連度を自動的に算出する手法では、コーパス内に出現する語しか扱えないという欠点がある。そのため、広範囲の語をカバーするためには、広範囲の内容をカバーするコーパスが必要となる。

近年では、より広範囲の語をカバーするために Web をコーパスとして用いることが提案されている。しかし Web 上の文書は莫大であり、直接収集し、解析するためには非常に大きな時間コストと設備コストがかかる。そのため、Web 全体での語の出現頻度や 2 語の共起頻度を獲得するためには従来のコーパスを用いたシソーラス構築とは異なる工夫が必要である。そのような工夫の一つとしてに Kilgarriff らは検索エンジンを用いた手法を紹介している [22]。「語 w_a」をクエリーとして検索エンジンを利用すると、語 w_a の Web 上でのヒット件数が得られる。検索エンジンは非常に多くのページをクローリングしているため、このヒット件数を語 w_a の Web 全体での出現頻度と近似できる。同様にして、「語 w_a and 語 w_b」をクエリーとすれば、Web 上での語 w_a と語 w_b の共起頻度を獲得することができる。

検索エンジンから獲得できる頻度情報を用いて関連度を算出する手法としては、次のようなものがある。Heylighen は検索エンジンのヒット件数を用いた語の関連度の尺度により、語の分類や語の曖昧性解消、より優れた検索エンジンの開発の可能性を示唆している [17]。Baroni や Tuerney は、類義語を同定するために、検索エンジンを用いた語の関連性の尺度を提案している [3, 51]。Turney はその結果を用いることで TOEFL のシソーラスの問題で平均的な学生よりもよい得点を挙げたことを報告している。佐々木らは検索エンジンの上位ページとヒット件数を利用した専門用語集の自動構築を行っている [59]。Szpektor は名詞ではなく動詞の関連度を検索エンジンを用いて定義している [50]。これら検索エンジンを用いて関連度の計算を行っている研究では、条件付き確率や表 3.8 の確率手法で定義されているような相互情報量、Jaccard 係数が計算指標として用いられている。

第4章　制約付きクラスタリングを用いた論文集合の構造化

4.1　はじめに

本章では、論文集合を各論文分野のカテゴリに分類することを考える。文書集合の分類については、今まで様々な文書分類や文書クラスタリングを初めとする様々な分類手法が提案されてきている。しかし、これらのような従来の分類手法が対象としてきた文書集合と比べ、論文集合には大きく異なる点がある。それは、分類すべきカテゴリが時間と共に変化していくという点である。例えば、近年の「人工知能」に関する分野であれば、以前は存在しなかった「Web マイニング」や「複雑ネットワーク」といった分野が出現している。

このように各カテゴリというのは時間を追って変化していくものである。1つのカテゴリは時間の経過共に成長したり、または縮小したりする。さらに大きくなったカテゴリが2つのカテゴリに分裂することもある。また新たなカテゴリが発生することもある。本章で、このような時間的に変化するカテゴリを考慮した論文分類を行うことを目的とする。しかし、既存手法を用いてこのような時系列変化を俯瞰的に捉えることは難しい。

そこで本研究では、従来のアプローチを組み合わせ、お互いの欠点を補完することで、論文分野の時間的な変化を捉えることのできる手法を提案する。特に、制約付きクラスタリングという新たなクラスタリング手法が大きな特徴である。

4.2　時系列的な観点から見た従来手法

ある論文集合を分類することを考えるとき、我々は２つの情報を考慮していると考えられる。一つは現在あるカテゴリの特徴、もう一つは分類対象となる文書集合の関係性である。分類対象をなるべく現在あるカテゴリに分類することを考えるが、その論文集合に偏りがある場合、新しいカテゴリを作ることもあり得る。例えば、「言語処理」「人工知能」「ネットワーク」というカテゴリがあり、分類対象の論文がそれぞれ３つの分野に均等に分かれるなら、そのままのカテゴリを用いるだろうし、もし論文集合の８割が「ネットワーク」に関係する論文であれば、カテゴリ「ネットワーク」を２つないし３つのカテゴリに分けると考えられる。

このように人間が捉えるカテゴリというものは時系列的に変化していくもので

ある。従来の文書を分類するアプローチを考えると、文書分類と文書クラスタリングの２つの手法があげられる。しかし、時系列的な分析という観点から考えた場合、これらの手法には大きな問題点がある。

4.2.1　時系列的な観点から見た文書分類

文書分類は、あらかじめ与えられたカテゴリに文書を分類する手法である。この手法では、人手によってカテゴリ分類された文書集合を訓練データとして、各カテゴリへの分類器を学習し、その分類器を用いてカテゴリへの分類を行う。これを論文集合にあてはめると、図 4.1 のようにあらかじめ分野カテゴリごとに分類された論文集合を用いて分類器を学習し、それを用いて新規論文を各カテゴリに分類すると考えられる。このプロセスを時系列的な観点から捉えれば、過去の時点のデータを用いて現在のデータを解析していると言える。過去の時点のデータとは、あらかじめ分類された論文集合であり、現在のデータとは新規論文である。言い換えれば、文書分類を用いた論文のカテゴリ分類では、現在のデータを考慮していないと考えられる。

このように論文のカテゴリ分類に文書分類の手法を用いる場合、過去の時点から見た現在の状況を反映できる反面、現在の論文集合のみで考えた状態を反映する事ができないのである。

4.2.2　時系列的な観点から見た文書クラスタリング

文書クラスタリングは、類似性や関連性といったある文書集合全体の関係性を手がかりとして、関係の強い論文が１つのグループ（クラスタ）にまとまるように、全体を分割する手法である。これを論文集合の分類にあてはめると、新規論文集合の中で関係の強い論文同士を一つのクラスタにまとめているといえる。時系列的な観点から捉えれば、現在の状況のみで論文を整理していると考えられる。そのため、過去の時点でのカテゴリを考慮することができないため、図 4.2 のように各年によって全く異なる結果が得られてしまう可能性が高い。

このように論文のカテゴリ分類に文書クラスタリングの手法を用いる場合、現在の論文集合のみで考えた状態過去の時点から見た現在の状況を反映できる反面、過去の時点から見た現在の状況を反映する事ができないのである。

4.3　提案手法の概要

4.3.1　基本的な考え方

本論文では、時系列的な変化を考慮した論文のカテゴリへの分類手法を提案する。

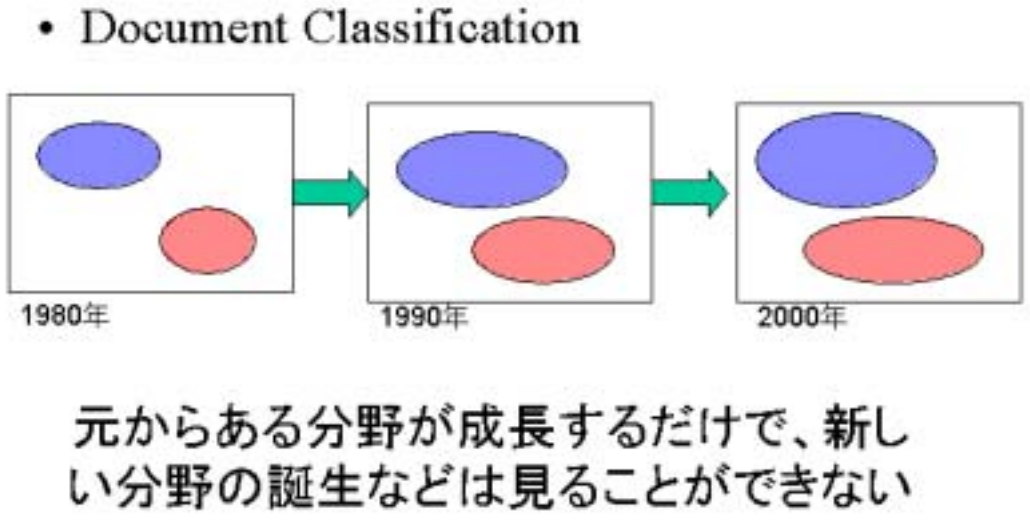

図 4.1: 文書分類の問題点

論文集合のカテゴリ分類を考えると、文書分類は、過去のデータを生かしているもの現在の論文集合の状態を反映していないことが問題である。また、文書クラスタリングでは、逆に現在のデータを生かしているものの、過去のデータを無視していることが問題点となっている。そこで、本研究では、2 つのデータ系列を足し合わせ、そのデータ上でクラスタリングを行うことで、両方の影響を加味したクラスタリングを行うことを目指す。

まず、ある論文集合 D があったとき、それらを既存のカテゴリに分類する。このカテゴリ分類を同カテゴリ行列 C によって表す。同カテゴリ行列 C とは、各論文を要素とする n^2 行列であり、2 つの論文が同じカテゴリに所属しているとき、その 2 つの論文にあたる要素が 1 となった行列である。次に、論文間の類似度に基づき類似度行列 S を構成する。S は各 2 論文の類似度を表す隣接行列である。

以上 2 つの行列を C, S とすると、図 4.3 のように C, S を合わせることで制約付きネットワークを構成する。制約付きネットワークとは、図 4.3 のようにこれは類似度の逆数を距離とする論文ネットワークに対し、さらに各 2 論文が同じカテゴリに含まれるか否かで、制約を付加したネットワークである。ここでは、類似度に対し、さらに同じカテゴリに含まれることを数値として付加する。

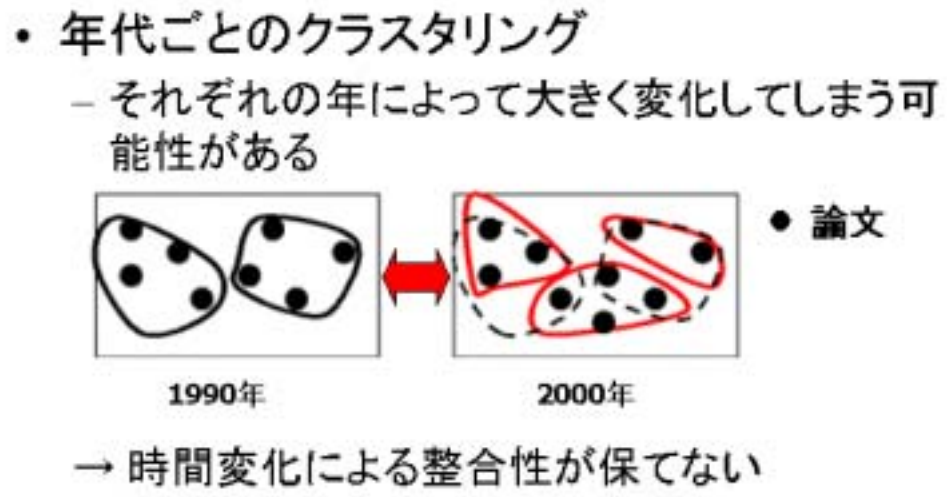

図 4.2: 文書クラスタリングの問題点

このとき制約付きネットワークの隣接行列 R は次式で表されるものと定義する。

$$R = (E - r)S + rC \tag{4.1}$$

E は単位行列を表す。r は S による制約の強さを表すパラメータであり、制約行列と呼ぶ。制約付きネットワークを構築することで、2 つの系統のネットワークの影響を考慮することができる。したがって、このような制約付きネットワーク上でクラスタリングを行うと、2 系統のデータの影響を受けたクラスタ結果が得られると考えられる。

4.3.2 提案手法の流れ

本研究では、類似度や引用関係によるネットワークに対して、カテゴリ分類によるネットワークによる制約をつけることで、カテゴリ分類による影響を加味した文書クラスタリングを行う。

手法の流れは以下に示すとおり。

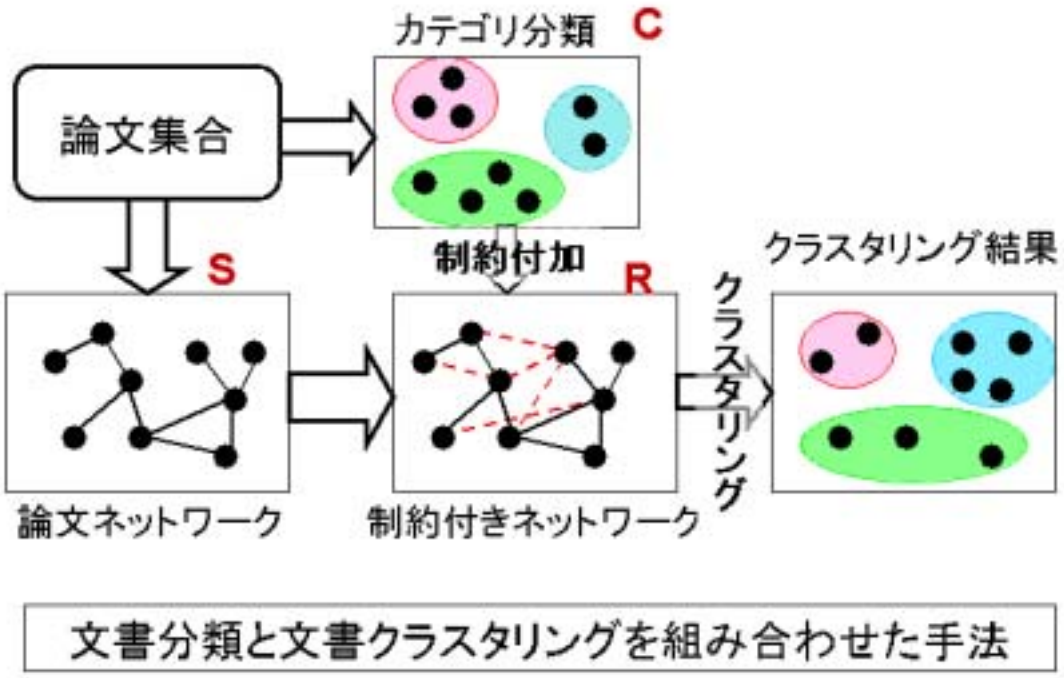

図 4.3: 制約付きクラスタリング

1. 類似度や引用関係などのデータから、論文ネットワークを構築する。
2. 次に何らかの手法で論文のカテゴリ分類を行う。
3. 同じカテゴリに分類された論文同士を結んだ、カテゴリネットワークを構築する。
4. 2 つのネットワークから、制約付きネットワークを構築する。
5. 制約付きネットワーク上でクラスタリングを行う。
6. 既存カテゴリとクラスタ結果の対応づけを行う。

4.4 提案手法の実装

4.4.1 論文ネットワークの構築

論文ネットワークを構築する場合、内容の類似性や引用関係、共著関係をなどを手がかりとするのが一般的である。しかし、引用関係や共著関係はあらかじめ書誌情報が論文に付随していることが前提であり、取得した全ての論文に対して適用できるわけではない。また、収集した論文に引用関係や共著関係がある確率は高くないため、ネットワーク自体がスパースになってしまう可能性が高い。

本論文では類似度を手がかりとしてネットワークを構築する。類似度の算出方法としては、最も一般的な手法の 1 つである、文書ベクトルを用いたベクトル空間法を用いる。ただし、論文の本文全てを用いる場合、文書ベクトルのノイズが多くなってしまうので、今回はアブストラクトを用いる。各論文のごとに、アブストラクトに出現する語の tfidf 値を要素とする文書ベクトルを定義する。式 (4.2) のような文書ベクトルの cosine 値を各論文間の類似度とする。

$$cosine = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} \tag{4.2}$$

そして、論文をノード、論文間の類似度をエッジの重みとして論文ネットワークを構築する。

4.4.2 既存カテゴリへの分類

論文を自動的にカテゴリ分類する手法は、今まで数多く研究されている。しかし、ほとんどの手法が人手で分類したデータを元に分類器を作成する教師付き機械学習を用いた手法であり、あくまで人手による分類が 100%の正解としている。

そこで本研究では、人手によるカテゴリ分類をそのまま用いる。人手による分類は、個々の研究者が行っている場合と、学会などがあらかじめ分類している場合があるが、いずれの場合も適用可能である。

また、このカテゴリ分類を用いてカテゴリネットワークを構築する。カテゴリネットワークとは、論文をノードとし、同じカテゴリに分類された論文同士にエッジを結んだネットワークである。今回は、モデルの簡単化のためにエッジに重みのない重み無しネットワークとする。

4.4.3 制約付きクラスタリング

類似度による論文ネットワークに対して、カテゴリ分類によるネットワークを用いて制約を付加することで、制約付きネットワークを構築する。類似度ネットワークの隣接行列を S、カテゴリネットワークの隣接行列を C とすると、制約付きネットワークの隣接行列を R は、式 4.1 より次式のようになる。

$$R = (1-r)S + rC \tag{4.3}$$

今回は、モデルの簡単化のために r を行列ではなく単なる定数とする。このように構築した制約付きネットワーク上でクラスタリングを行う。クラスタリング手法としては、Newman 法を用いる。

Newman 法は、階層的クラスタリング手法の一つであるが、クラスタリングを評価関数 Q の最大値導出問題に置き換えた手法である [36]。評価関数 Q とは、各クラスタの結合度を表す関数であり、Q が大きいほど各クラスタ内の結合が強いことを表している。Newman 法では、Q の高い状態がより適切にクラスタリングされた状態であると定義している。そして、Q の最大値を求めることで、そのネットワークに最適なクラスタリング結果を得ることを目標としている。

評価関数 Q は次式で表される。

$$Q = \frac{1}{2m}\left[\left(\sum_{v,w} A_{vw}\delta(c_v, c_w)\right) - \left(\sum_{x,w} \frac{k_v k_w}{2m}\delta(c_v, c_w)\right)\right] \tag{4.4}$$

k_v は頂点 v が持っているエッジの本数、m は全エッジ本数の合計、c_v は頂点 v が属しているクラスタを表している。$\delta(c_v, c_w)$ はクロネッカーの δ である。式 (4.4) の第 1 項において、A_{vw} は頂点 v, w 間のエッジの有無を表しており、また頂点 v, w が同じクラスタのときのみ、$\delta(c_v, c_w) = 1$ となる。つまり、第 1 項は各クラスタ内に含まれるエッジの本数の合計を表している。同様に第 2 項においては、$\frac{k_v k_w}{2m}$ は頂点 v, w 間にエッジが引かれる確率を表しているため、第 2 項は、各クラスタ内に含まれるエッジの本数の合計の期待値を表している。

すなわち、評価関数 Q とは、クラスター内に存在するエッジの本数の合計が期待値からどの程度ずれているかを相対的に表した値である。クラスター内のエッジ本数の和が期待値と同じなら $Q = 0$、それより強いクラスターなら $Q > 0$ であり、弱いクラスターなら $Q < 0$ となる。Q が最大であるとき、各クラスター内での結合度が最大であるので、ネットワーク全体として最も良くクラスタリングされた状態であると考えられる。

しかし Q の最大値を求める場合、エッジ数 m、ノード数 n のとき、計算量が $O(n^3)$ もしくは $O(m^2 n)$ となり、大きくなってしまう。そこで Newman 法では Greedy アルゴリズムを用いて Q の値が極大値をとるようにクラスタリングを行う。すなわち、Q の変化量 ΔQ が最大になるようなクラスタのマージを繰り返していくことで、Q の極大値を求める。そして、Q が極大値となった時点でクラスタリングを終了する。

Newman 法と betweenness クラスタリングを比較すると、Newman らにより Newman 法は betweenness クラスタリングとほぼ同じ精度のクラスタリング結果が得られることが示されている。また、Newman 法の時間計算量は $O((m+n)n)$ もしくは $O(n^2)$ であり、時間計算量が $O(m^2 n)$ あるいは $O(n^3)$ である betweenness クラス

タリングと比べ、計算量が少なく、高速な手法となっている。そのため、Newman 法はノード数やエッジ数が大きい大規模ネットワークに適用可能である。

ここでは、式を次式のように書き換えた重み付き Newman 法を提案する。

$$\begin{aligned} \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) \\ e_{ij} &= \frac{\text{クラスタ } i, j \text{ 間のエッジの重みの和}}{\text{全エッジの重みの和}} \\ a_i &= \frac{\text{クラスタ } i \text{ 内の頂点と結び付いたエッジの重みの和}}{\text{全エッジの重みの和}} \end{aligned} \tag{4.5}$$

これにより、制約付きネットワークのエッジの重みの違いを考慮したクラスタリング結果が得られる。

4.4.4 カテゴリの同定

制約付きクラスタリングによって得られたクラスタと既存のカテゴリを対応づける必要がある。カテゴリの同定は、多数決法により行う。多数決法とは、制約クラスタリング結果から見て共通する論文数がもっとも多い既存カテゴリを対応カテゴリとする手法である。この場合、クラスタ結果と既存カテゴリが多対 1 対応となる可能性がある。

以下、1 つの既存カテゴリ C_i に対応するクラスタ数によってカテゴリの変化パターンを定義する。

成長 カテゴリ C_i に対応するクラスタが 1 つできた場合、カテゴリ C_i は特に変化することなく、カテゴリ C_i に含まれる論文が増えている。そのため、これをカテゴリ C_i の成長と定義する。

分裂 制約付きクラスタリングの結果、カテゴリ C に対応するクラスタが 2 つ以上できた場合、カテゴリ C_i が 2 つに分裂したものと定義する。

発生 カテゴリ C_i が分裂した際、最もサイズの大きいカテゴリを C_i とみなし、それ以外のカテゴリは新たに発生したカテゴリと定義する。

4.5 評価実験

本節では提案手法の評価を行う。

ある論文集合のカテゴリ分類の評価を行う場合、人間によるカテゴリ分類を正解として適合率、再現率を測る評価手法が一般的である。しかし、カテゴリの分裂や発生といった現象を考慮したデータがあるわけではないので、直接それらを評

価することは難しい。特に新しく収集した論文をカテゴリ分類したとしても、やはり正解データが無いために評価が難しい。

そこで本論文では、ある学会や論文誌の論文集合の過去のカテゴリ分類の変化を正解データとし、それらをどの程度再現できるか、ということで提案手法の評価を行う。

4.5.1 評価実験の概要

本論文では、ある過去の 1 時点のカテゴリ分類をもとに制約付きクラスタリングを行い、それによりその時点からみて未来のカテゴリ分類が再現できるかどうか、ということで評価を行う。

具体例をもって説明しよう。ある学会において、1980 年では、その年の論文が、図 4.4 のように「言語処理」「人工知能」の 2 つのカテゴリに分類されているとする。また、1990 年では、その年の論文が、図のように「言語処理」が分裂してできた「形態素解析」「機械翻訳」「意味解析」の 3 つのカテゴリと、「人工知能」が分裂してできた「学習」「エージェント」2 つのカテゴリにそれぞれ分裂しているものとする。このようなデータがあるとき、まず、1990 年の論文とそのカテゴリ分類を正解データとする。次に 1990 年の論文を、1980 年のカテゴリ分類、つまり「言語処理」と「人工知能」に再分類する。言い換えれば、「形態素解析」「機械翻訳」「意味解析」の 3 つのカテゴリと、「学習」「エージェント」の 2 つのカテゴリをそれぞれ強引に 1 つのカテゴリとみなす。そして、この 1980 年のカテゴリ分類によって制約を付加した制約クラスタリングを行う。その結果、図のように 1990 年のカテゴリ分類と一致したクラスタ結果となれば、提案手法は、有効に働いていると言える。

このように過去の 1 時点のカテゴリ分類用いて、それより進んだ過去の時点のカテゴリ分類をどの程度再現できるかどうかで、提案手法の評価を行う。そして、提案手法の効果を確認するために、文書クラスタリングのみの場合 (r=0) と文書分類のみの場合 (r=1) と提案手法を比較する。

4.5.2 評価に用いる指標

提案手法の数値的な評価は、適合率と再現率、およびクラスタ数のズレによって測る。

どちらの指標も、出力されたクラスタに、対応するカテゴリの語がどの程度の割合で含まれているかを表す。適合率と再現率は次式で表される。

$$\text{適合率} = \frac{\text{正解論文数}}{\text{出力クラスタサイズ}} \text{再現率} = \frac{\text{正解論文数}}{\text{対応カテゴリサイズ}} \tag{4.6}$$

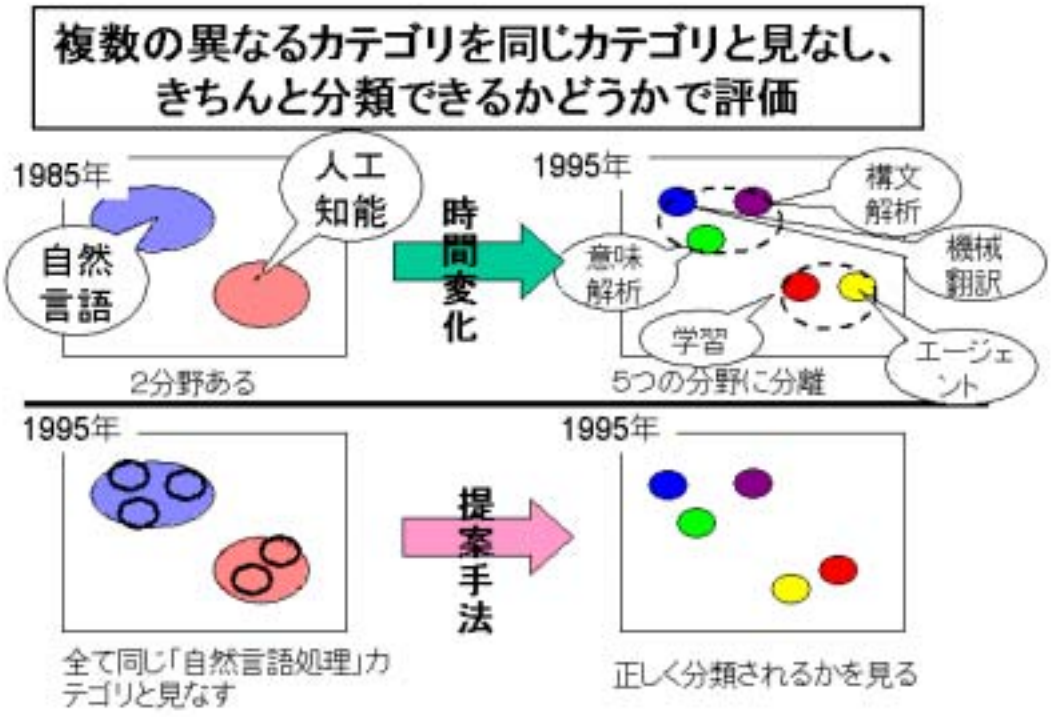

図 4.4: 制約付きクラスタリングの評価実験手法

4.5.3 カテゴリの再分類

今回用いる arXiv.org には、元々36 個のカテゴリが存在している。本実験では、それらを 8 つのカテゴリに再分類した。この 8 つのカテゴリは、1993 年から 1996 年の論文で出現回数の多いカテゴリから上位 8 つを取り出したものである。

各論文には主たる分野の他に、関連する分野がいくつか割り当てられる。この時、一つの論文に割り当てられている分野同士を分野共起していると呼び、またこの分野共起している頻度を分野共起頻度と呼ぶ。

本来ある 36 個のカテゴリの 8 つのカテゴリへの再分類は、この分野共起頻度を用いた。つまり各カテゴリは、8 つのカテゴリのうち、最も分野共起頻度の高いものに再分類を行う。カテゴリの再分類を図 4.5 に示す。

Networking and Internet Architecture	Artificial Intelligence
Networking and Internet Architecture	Artificial Intelligence
Operating Systems	Computer Vision and Pattern Recognition
Other	Databases
Performance	Information Theory
	Learning
Data Structures and Algorithms	Multiagent Systems
Data Structures and Algorithms	Neural and Evolutionary Computing
Discrete Mathematics	Robotics
Distributed, Parallel, and Cluster Computing	
Numerical Analysis	Computational Complexity
	Computational Complexity
Digital Libraries	Architecture
Digital Libraries	Computational Engineering, Finance, and Science
Computers and Society	Computer Science and Game Theory
Multimedia	Cryptography and Security
	General Literature
Computational Geometry	Mathematical Software
Computational Geometry	
Graphics	Logic in Computer Science
	Logic in Computer Science
Computation and Language	Programming Languages
Computation and Language	Software Engineering
Human-Computer Interaction	Symbolic Computation
Information Retrieval	
Sound	

図 4.5: カテゴリの再分類

4.5.4 評価実験とその結果

本評価実験では、あらかじめカテゴリ分類された論文データを使用する。今回は、Cornell 大学の arXiv.org[1] から Computer Science の分野の論文のうち、1993 年から 2005 年までの 13 年分収集し、それを実験データとした。

また、arXiv.org では元々36 個のカテゴリーがある。それを本実験のためにそれらのカテゴリーを表のように 5 つのカテゴリーに再分類した。本実験の手順を以下に示す。

1. 論文集合を発行年ごとに分類し、それぞれ y 年の論文集合を P_y とする。。
2. 論文集合 P_y の各論文の類似度を計算し、類似度行列 S_{P_y} を算出する。
3. 表の 5 つのカテゴリ分類を基に P_y の各論文をカテゴリ分類し、カテゴリ行列 C_{P_y} を算出する。

[1] http://arxiv.org

4. S_{P_y}, C_{P_y} を用いて、式 4.7 から制約付きネットワーク行列 R_{P_y} を算出する。

5. R_{P_y} をもとに制約付きクラスタリングを行う。

6. 図 4.5 のカテゴリ分類に基づいて、多数決法によりクラスタ結果と 36 個のカテゴリを対応させる。

7. 適合率・再現率を計算する。

8. 1.〜7. を $P_y(y = 1993 \cdots\cdots 2005)$ について行う。

$$R_{P_y} = (1 - r)S_{P_y} + rC_{P_y} \tag{4.7}$$

まず、予備実験として r を色々と変化させ、最も値の良い r を選択する。r を 0.001〜1 まで変化させた場合の各 P_y の適合率・再現率をあらわしたグラフを図 4.6 に示す。図 4.6 において、縦軸は割合を横軸は r の値を表す。

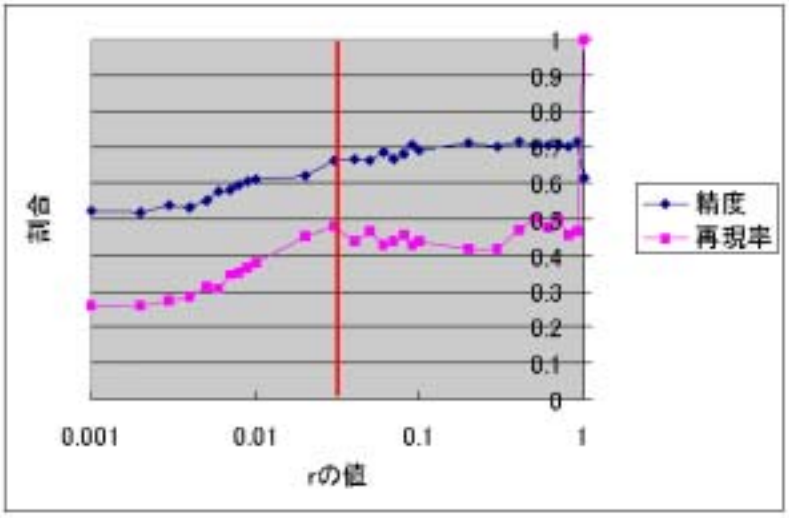

図 4.6: 論文カテゴリ分類の予備実験結果

図 4.6 において、精度は $r = 0.1$ 前後までは単調増加し、$r \geq 0.1$ ではあまり変化していない。また、再現率は $r = 0.03$ 前後までは単調増加し、その後 $0.03 \leq r \leq 0.3$ では減少し、再び $r \geq 0.3$ において増加している。

数値的には、$r = 0.9$ 付近が精度・再現率共に高くなっている。しかし、$0.3 \leq r \leq 1$ において再現率が高いのは、大きなクラスタができているためである。一つのク

ラスタが大きければ、多数の論文が含まれるため、再現率は高くなる。しかし、このようにクラスタが大きく、クラスタ数が少ない場合は本手法に置いては望ましくないと考えられる。そこで、再現率が高く、また精度も $r = 0.3$ 付近とほとんど変わらない $r = 0.03$ を今回で最も良い値と考えられる。

次に r=0.03 と、r=0 つまり文書クラスタリングのみの場合と、r=1 つまり文書分類のみの場合とをそれぞれ比較する。比較結果を表 4.1 に示す。

表 4.1: 従来手法との比較

	文書クラスタリング	提案手法	文書分類
精度	0.520	0.662	0.616
再現率	0.263	0.480	1.00

まず、提案手法 (r=0) と文書クラスタリング (r=0.03) を比較すると、精度、再現率共に提案手法の方が高い。また、提案手法と文書分類 (r=1.0) を比較すると、精度は提案手法が上回っているもののの、再現率は文書分類の手法が高い。これは、文書分類のデータ用いて論文の分類を用いているため、当然である。しかし、実際には再現率が高い文書クラスタリングにおいては各クラスタのサイズが大きいために、本手法で求める結果としてはあまり良い結果ではない。

これより、提案手法を用いることで文書クラスタリング、文書分類よりも高い精度で論文を分類することができる。以上より、提案手法の有効性を確かめることができた。

4.6 議論

本章では、制約付きクラスタリングにより、時系列的なカテゴリの変化を考慮した論文のカテゴリ分類を行った。評価実験にでは、単なる文書クラスタリングの手法と文書分類の手法と提案手法を比較することで、提案手法が確かに論文の分類において有効であることを示した。

今回用いている過去のデータというのは 1993 年〜1996 年で最も多く出現した 8 つのカテゴリである。しかし、本来ならば複数の時点での過去のデータを用いるべきである。過去のデータをきちんと連続的に用いることで、より精度・再現率が向上すると考えられる。

また、カテゴリの実験の結果はカテゴリの再分類にも影響されるため、より正確なカテゴリの再分類を行う必要がある。これについては、論文検索サイトにおいて論文にインデックス付けするためのより詳細なカテゴリの分類が、いくつものサイトに存在しているため、これらを適用することで、さらに適切にデータを取り扱うことを考えたい。

本実験では、予備実験により最もよい値を返す r を決定している。しかし、制約係数 r、つまり過去のデータをどの程度分類結果に反映させるかは、本来自動で決定されるべきである。今後は、そのように r を自動決定する手法についても検討していきたい。

4.7 関連研究と位置づけ

近年、コンピュータの普及に伴い、電子化された文書が爆発的に増え続けている。そのためそれらの文書を分類し、整理する研究は数多く行われている。特にWeb 上の文書など、無秩序に存在している文書を整理することは、学術的にもビジネス的にも大きな意味を持ってくる。

文書を整理する技術は大きく分けて、文書分類と文書クラスタリングに大別できる。本節では、この 2 つの技術について紹介し、その問題点について述べる。

4.7.1 文書分類

文書分類は、ある論文集合を与えられた分野・カテゴリに分類する作業であり、論文の整理から Web 文書の構造化、メールの自動分類、スパムフィルタリングなど様々な分野に適用されている。特にここ 10 年で大きく進歩している。

文書分類の手法は大きく分けて 2 つに分けられる。1 つは、人手による分類手法である。これは与えられた文書集合に対して、人が各文書に最も適当なカテゴリを判断し、分類していく手法である。主に、Yahoo カテゴリ[2]や OpenDirectoryProject(ODP)[3], Looksmart[4]など、Web ディレクトリの作成に用いられている。人が行うので、文書分類の手法では最も精度が高いと考えられる。しかし、分類する人の判断・知識に分類結果が大きく影響されるため、分類の一貫性が低くなる可能性がある。ODP などでは、複数の人間によって判断を行うことで、その問題を回避している。しかし、人手による分類は非常にコストがかかるため、それほど多く行われているわけではない。

もう 1 つは計算機による自動分類である。これは、クラス分類のアプローチを用いて自動的に文書分類を行う分類器を作成する手法である。以前は、このような分類器は人手で作られていた。しかし、高精度の分類器を人手で作り、かつ維持していくのは非常にコストがかかる。近年では機械学習の手法を用いて分類器を自動作成する手法が一般的である。これは、あらかじめ人手で分類された事例データを元に分類器を学習する教師付き学習の手法である。Sebastiani らによって分類器の学習に用いられる一般的なアプローチ、決定木、NaiveBayes 手法、kNN

[2]http://dir.yahoo.co.jp/
[3]http://dmoz.org/
[4]http://search.looksmart.com/

法、SVM などが紹介されている [46]。

決定木では、カテゴリとその事例データおよび属性と値から分類規則を生成し、それを木構造で表現する。決定木では、葉はカテゴリ、葉以外のノードは属性を表し、その各ノードの枝はその属性の属性値に対応している。、この決定木上で各属性の属性値に対応する枝たどっていくことで、文書分類を行うことができる。学習データから決定木を構築する方法は複数有り、文書分類の分類器の学習に用いられる物としては、ID3[13]、C4.5[6, 28]、C5[31] などが有名である。決定木による分類器は、属性数が多くなると過学習してしまう反面、分類過程がわかりやすいという利点がある。

Naive Bayes 分類は、観測データを用いて各カテゴリに分類される確率を求める学習法であり、Koller ら [24] や Larkey ら [26] が用いている。次式のような確率分布の和を元に、各カテゴリへの分類確率を計算し、最も確率の高いカテゴリへ分類を行う。式 4.8 では、d_j が文書 j の文書ベクトル、c_i がカテゴリ i w_{kj} が文書 j に含まれる k 番目の語を表している。つまり、文書 j がカテゴリ i に分類される確率 $P(d_j|c_i)$ は、文書 j に出現する語 w_{kj} がカテゴリ i に含まれる確率 $P(w_{kj}|c_i)$ で表されている。NaiveBayes は確率的にカテゴリへの分類を確率的に表せる反面、計算コストが大きくなりという欠点がある。

$$P(d_j|c_i) \approx \prod_{w_{kj} \in d_j} P(w_{kj}|c_i) \times \prod_{w_{kj} \notin d_j} (1 - P(w_{kj}|c_i)) \tag{4.8}$$

kNearestNeighbour による分類は、定義した文書間の距離を元にその文書に最も近い k 本の論文を取得し、それらの中で最も多いカテゴリに分類する手法である。Yang ら [55] や Lewis ら [29] が文書分類に適用している。この手法では、属性の選択が不要、大規模なデータにも適用可能という利点があるが、距離の定義やわずかな違い大きく影響されるという欠点がある。

Support Vector Maching は 2 クラスの分類を行う学習機械の一種である。SVM では、n 次元空間に学習点を配置し、正の訓練点と負の訓練点との境界距離が最大化するように分離超平面を構築する。この分離超平面を元に正負の分類を行うのである。SVM は，特にパターン認識の能力において、最も優秀な学習モデルの 1 つであることが知られており、文書分類の分野でも近年最も多く用いられている手法である [20, 10, 11]。

しかし、いずれの手法もあらかじめ定められたカテゴリに分類するということを目的としているため、カテゴリ自体が変化する場合などに対応することはできない。また、文書クラスタリングと比べて教師付き学習手法であるため、精度を上げるためには大量の事例データが必要となる。

4.7.2 文書クラスタリング

文書クラスタリングは、文書間の類似度を定義し、それをもとに文書をクラスタリングする手法であり、検索結果の表示の整理や、検索結果の精度・スピード向上など、情報検索の分野で主に用いられている [54, 1]。

従来のクラスタリング手法

クラスタリングとは類似したものをクラスタとしてまとめあげる技術であり、学習データのいらない教師無し学習手法である。クラスタリング手法は大きく分けて 2 つに分けられる [21]。

一つは、k Means 法をはじめとする分割最適化手法である。これはデータの分割の良さの評価関数を定義し、その評価関数を最適化する分割を探索する手法である。k-Means 法では、セントロイド c_i(クラスタの重心点) をクラスタの代表点とし、次式のような評価関数を最小化することで、k 個に分割する際の最適なクラスタリング結果を求めている。

$$\sum_{i=1}^{k} \sum_{x \in C_i} (D(x, c_i)) \tag{4.9}$$

上式で x は各データ、$D(x, c_i)$ はデータ x と重心 c_i の距離を表している。分割最適化手法では、最小解の探索の計算コストが高くなることが知られている。

もう一つは階層的クラスタリングの手法である。これは、各クラスタを順にマージまたは分割していくことでクラスタリングを行う手法である。この手法は大きく分けて凝集型と分枝型の 2 つに分けることができるが、文書クラスタリングでは、主に凝集型クラスタリング (agglomerative clustering) が用いられる。これは、N 個のデータが与えられたとき、初期状態を各データ 1 個で 1 クラスタと見なす。そして定義した距離関数をもとに、これらのクラスタを最も距離の近い順にクラスタをマージしていく手法である。この手法に用いられる距離関数は表 4.2 のようなものがあげられる。

表 4.2: 階層的クラスタリングで用いられる距離関数 $D(c_i, c_j)$

手法	最大距離法	最小距離法	群平均法
$D(c_i, c_j)$	$\max_{w_k \in c_i, w_l \in c_j} Sim(w_k, w_l)$	$\min_{w_k \in c_i, w_l \in c_j} Sim(w_k, w_l)$	$\frac{1}{n_i n_j} \sum_{w_k \in c_i} \sum_{w_l \in c_j} Sim(w_k, w_l)$

近年のクラスタリング手法

近年では、データをネットワークで表した上でデータを分析する手法が着目されており、ネットワーク上でのクラスタリングも数多く提案されている。例えば

betweenness クラスタリングは、グラフ[5]の betweenness というエッジの媒介性を表す指標（あるエッジが他のエッジの最短パスにどの程度の割合で含まれているか) に注目し、できるだけ部分グラフをつなぐような betweenness の高いエッジを削除していくことにより、密度の濃いサブグラフを同定する手法である [14]。また、IsingModel クラスタリングは、ネットワークのクラスタリングの状態を Ising Model と呼ばれる量子力学のエネルギー関数で近似し、エネルギー関数の最小値化問題の手法を用いて、最適なクラスタリング状態を決定する手法である [48]。これらは語の関係分析にも用いられており [53, 35, 39]、今後文書クラスタリングの分野にも用いられることが考えられる。

しかし、これらようなクラスタリング手法は、データ集合に対して類似したものをまとめることはできるが、その結果に何らかの意味を与えることは難しい。例えば、クラスタリング結果にラベル付けをする手法などがあるが、高い精度が得られないのが現状である。

教師ありクラスタリング手法

教師ありクラスタリングとは、神嶌らが提案している手法であり、事例データの集合から目標とする分割に望ましい規準を獲得し、それをもとに未知のデータ集合のクラスタリングを行う手法である [61]。この手法はクラス分類とクラスタリング手法を合わせているという点で、提案手法と非常に似通っている。

しかし、この手法が目的としているのは、目標とする分割は明確だが、その分割規準が明確でない場合に、自動的に分割を導くクラスタリングがうまく機能しない、という問題点を解決することである。そのため、目標分割を導く規準を事例データから学習し、その規準を用いてクラスタリングを行っている。つまり、事例データからの学習という観点からクラス分類をクラスタリングに組み合わせているのである。

一方、提案手法では、時系列的に変化していくデータ集合に対して、ある 2 つの連続した時点、例えば 1980 年と 1981 年とで、クラスタリングが全く異なる分割結果を返す、という問題点を解決することである。そのため、過去の 1 時点での分割結果による影響を与えた上で、クラスタリングを行っている。つまり、時系列的な影響力という観点からクラス分類をクラスタリングと組み合わせている。つまり、教師有りクラスタリングとは観点が大きく異なるのである。

[5]ネットワークは、エッジに重みや長さなどの数値が付加されているのに対し、グラフはエッジに数値の付加されていない、接続関係だけを表すものである。

第5章　結論

本論文では、既存のサーベイ支援システムを紹介し、それらに必要と思われる2点の機能について提案を行い、その有効性を示した。

5.1　Web上の情報を用いた検索クエリ拡張について

本論文では、Web上から論文検索を支援するために、自動構築したシソーラスを用いた検索クエリ拡張の手法を提案した。提案手法では、検索エンジンを通してWeb上の知識を利用することで、・関連語の取得、・関連度の算出というシソーラス構築に必要な2つのプロセスを行っている。特に関連度の算出に関しては従来手法と同程度の精度を保っており、また従来手法と比べて、どのような語でも扱うことができることが大きな特徴である。

特に今回提案したシソーラスの構築手法は他の分野にも応用することができる。機械翻訳や文脈解析、省略解析などの様々な言語処理の研究においてシソーラスは重要なデータソースである。逆に言えば、既存のシソーラスがネックとなって、解決できていない問題などもある。そのような意味で、どのような分野・範囲の単語でも扱える今回のシソーラス構築手法は今後ますます重要な意味を持ってくるのではないかと考えられる。

近年ではWebは重要な言語処理として考えられており、その情報をいかに効率的に利用して言語処理に適用するか、といった研究が数多く行われている。それらの研究では、検索エンジンの利用や大規模なデータの処理への対応など、Web上の情報を用いるならではのアルゴリズムの工夫が必要となる。そのような一つのアプローチとして、検索エンジンを用いるアプローチが用いられているが、その有効性を疑問視する声もある。本論文では、そのような研究の一つとして検索エンジンを通してWeb上の情報を利用することの有効性を示せたと言える。今後、検索エンジンを利用した言語処理の可能性をさらに追求していきたい。

5.2 制約付きクラスタリングを用いた論文のカテゴリ分類について

本論文の後半では、、文書分類を文書クラスタリングの手法を組み合わせた制約付きクラスタリングによって、論文集合の構造化を行う手法を提案した。本論文では、時間的に変化していく論文カテゴリへのカテゴリ分類に既存の文書分類や文書クラスタリングが適用できないことを指摘した。それらの 2 つのアプローチの欠点を補完するために制約付きクラスタリング手法を提案した。実験結果では、文書分類および文書クラスタリングよりもよいカテゴリ分類の結果が得られており、本手法の有効性が示せていると言える。このように分類すべきカテゴリ自体が変化するという条件において、既存手法よりも高い精度分類結果が得られるのが大きな特徴である。

実際に人間が文書のカテゴリ分類を行う際に、判断基準に用いられるのは、論文の類似度だけではなく、その論文の引用文献・非引用文献の一致や、著者が共通している点、などが用いられる。今回は類似度のみのネットワークを用いているため、不十分と考えられるかもしれない。しかし、類似度のみを用いた場合でも従来手法より高い精度が得られており、手法の有効性を示すには十分と考えられる。今後は、文書間類似度だけではなく、引用文献・被引用文献情報や著者情報などの手がかりを用いて、より精度の高い結果を得ることを目標とする。

また、制約付きクラスタリングは様々な分野に応用できることが考えられる。本論文では、論文を対象としているため、類似のネットワークと人手によるカテゴリ分類を用いた。しかし、分類やクラスタリングが対象としているのは文書だけではなく、単語分類から画像認識まで考えられる。本手法は、分類とクラスタリングが考えられる対象では、どんな対象にも応用することできる汎用的な手法であるといえる。今後は、この手法を用いて他の対象を解析していくことも考えていきたい。

付 録A　マーケティングビジネスへの利用

ここでは、株式会社 HottoLink Inc. が、本論文の第 3 章の技術などを用いて開発した BlogWatcher Enterprise β について紹介する。

近年、Web 上ではブログや価格系比較サイト等の登場でネットの口コミ情報が注目を集めている。現在の日本では、1 分間に 200 件、1 日に換算すると約 300,000 件のブログの投稿があると言われており、1 年後には約 100,000,000 件のブログが出現すると言われている。BlogWatcher Enterprise は、世の中のブログや Web 日記から評判情報を自動抽出することができる最新技術を活かした製品であり、ブログ内でのキーワードの盛り上がり度（バースト度)、評判分析を行うことが可能となっている。

BlogWathcer Enterprise では、本研究の第三章の検索エンジンを用いたシソーラス構築の技術をブログ検索エンジンに用いることで、ある特定のキーワードとブログ内で関連が強い語を取り出す、キーワードマイニングを実現している。。図 A.1 のように関連する語句を取り出し、それを自動的にグループ化することで、一人でもブレインストーミングが行えるようになっている。

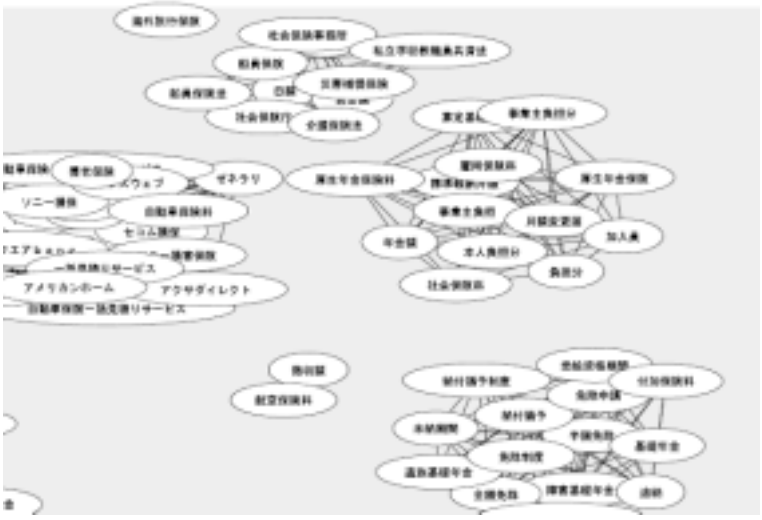

図 A.1: ある特定のキーワードの関連語句抽出と自動グループ化

謝辞

本研究を行うにあたり、非常に多くの方々のご指導・ご鞭撻を賜りましたことを、この場をお借りして心より感謝いたします.

指導教官である石塚満教授には、お忙しい身でありながら貴重な時間を割いていただき、日頃から多くのご指導・ご鞭撻を賜り、また本研究について指針を与えていただき、心より感謝いたします。

石塚研究室秘書の藤田メイコさんには、平時より楽しく快適に研究を行う環境を整えていただき感謝しています。

助手の土肥浩氏には、ミーティングなどで、ご指導やアドバイスを賜りましたこと、また、本業の傍ら研究室のサーバーやネットワークの管理・監視、スパムメールの排除まで行っていただき、快適な研究環境の保持に勤めてくださったことを心より感謝いたします。

石塚研究室の OB であり、現在、産業技術総合研究所譲歩技術研究部門ならびにスタンフォード客員研究員の松尾豊氏には、お忙しい中、毎週研究室に来ていただき、個々の研究へのアドバイスから、研究に対する心構えまで色々とご指導くださって感謝しております。また、単なる研究への指針・アドバイスにとどまらず、各論文の手直しや書き方の指導、研究者としてのあるべき姿や意義部会研究のあり方など、実に多岐に渡ってご教授いただいたと思います。また様々な方へ紹介していただき、修士課程ながら研究の面白さや醍醐味を味わえたことは非常に有意義でした。解くに学術研究のビジネス化に触れられたことは、自分を成長させる意味でも非常に意義深かったと思います。

国立情報学研究所の武田英明教授、ならびに市瀬龍太郎氏には、本研究にあたり多くのご指導・ご鞭撻をいただきました。おかげで、異なる観点から研究を捉えられたことができ、研究者としての視野を拡げることができたと思います。

株式会社のホットリンク方々には、社長の内山幸樹氏や R&D 事業部の末永氏、下大園氏には研究のビジネス化という点で、研究者と異なる点から様々なアドバイスをいただきました。実際のビジネスミーティングの場にも立ち合うことができ、また様々な方々とお会いすることができ、自分を磨くうえで大きなプラスにすることができました。

国立情報学研究所の大向一輝氏には、ホットラボにて研究のビジネス化や研究のまとめ方・進め方など様々な点においてアドバイスをいただきました。特に修士の研究をきちんとした形でまとめることができたのは、大向さんのご指導のたまものです。この場を借りてお礼申し上げます。

研究室の博士２年の森純一郎氏には、有益な指摘やディスカッションをしていただきました。また研究室内でのミーティングを取り仕切ってくださったおかげで、スムーズに研究を進めることができ、感謝しております。

博士２年で、現在留学中の岡崎氏には研究のみならず実生活の点までアドバイスをいただきました。おかげで、あまり悩むことなく研究を進めることができ、感謝しております。

研究室のOBである浅田氏、谷口氏、藤村氏とは、修士１年のみの間ではありましたが、研究に対して的確なアドバイスをいただいただけでなく、共にフットサルをして汗を流したり、また研究の合間の雑談などをしておかげで、非常に楽しく研究室での生活を送ることができました。この場を借りてお礼を申し上げます。

同じ学年である、金英子さん、櫛田氏とはお互いに刺激しあい切磋琢磨して研究を進めることができ、大変感謝しております。

また環境海洋学部の内田氏、柴田氏とは、４ヶ月にわたり共に輪読会を開催してきましたが、おかげで、自らの見識を深めると同時に研究をする上で非常によい刺激なったと思います。

東京理科大学の岡田、古川両氏とは研究興味が似ていることもあり、お互いに切磋琢磨しながら非常に楽しく研究を進められたと思います。

上記の方以外にも、日頃から多岐に渡りご指導くださったその他の石塚研究室の先輩・後輩諸氏に感謝いたします。

最後に、修士研究をこの石塚研究室で行うことができたことは、自分にとって非常に幸運なことだったと思います。非常に良い方々に囲まれ、２年間の有意義な研究生活を送れたことは今後の人生の糧としていきたいと思います。本当にありがとうございました。

発表文献

- 榊剛史，松尾豊，市瀬龍太郎, 武田英明, 石塚満，“論文データベースからの研究トピック抽出，” 第 19 回人工知能学会全国大会論文集 (CD-ROM)，(2005.6).
- 榊剛史，松尾豊，石塚満，“制約付きクラスタリングを用いた論文分類，” 第 19 回人工知能学会全国大会論文集 (CD-ROM)，(2006.6),(予定)
- 榊剛史, 松尾豊, 内山幸樹, 石塚満, “検索エンジンを用いた連想語の抽出，およびそのクラスタリング”, 第 5 回 Web インテリジェンスとインタラクション研究会,(2006.3), (予定)
- 榊剛史, 松尾豊, 石塚満, “時系列変化を考慮した論文ネットワークの解析”, 情報処理学会 第 68 回全国大会, (2006.3), (予定)

参考文献

[1] *A comparison of document clustering techniques*, 2000.

[2] D. Baker and A McCallum. Distributional clustering for text classification. In *Proceedings of SIGIR-98,21st ACM International Conference on Research and Development in Information Retrieval*, pp. 96–103, 1998.

[3] M. Baroni and S. Bisi. Using cooccurrence statistics and the web to discover synonyms in a technical language. In *In Proceedings of LREC2004*, pp. 26–28, 2004.

[4] P. Brown, V. Pietra, P. deSouza, J. Lai, and R. Mercer. Class-based n-gram model of natural language. *Comput. Linguist.*, Vol. 18, No. 4, pp. 467–479, 1992.

[5] W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, Vol. 16, No. 1, 1990.

[6] W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pp. 307–315. ACM Press, New York, US, 1996.

[7] J. Curran. Ensemble methods for automatic thesaurus extraction. In *Proceedings of the 2002 Conference on Empirical Methods in NLP*, pp. 222–229, 2002.

[8] J. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop of the ACL SIGLEX*, pp. 59–66, 2002.

[9] S. Dhillon. Enhanced word clustering for hierarchical text classification. In *Proceedings of the 8th ACM SIGKDD*, pp. 191–200, 2002.

[10] D. DRUCKER, V. VAPNIK, and D. WU. Automatic text categorization and its applications to text retrieval. In *IEEE Transaction Neural Networks*, pp. 1048–1054, 1999.

[11] S. Dumais and H. Chen. Hierarchical classification of web content. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 256–263. ACM Press, 2000.

[12] G. Flake, E. Glover, S.Lawrence, and C. Giles. Extracting query modifications from nonlinear svms. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pp. 317–324. ACM Press, 2002.

[13] N. Fuhr, S. Hartmann, G. Knorz, G. Lustig, M. Schwantner, and K. Tzeras. Air/x – a rule-based multistage indexing system for large subject fields. In *Proceedings of RIAO-91, 3rd International Conference "Recherche d'Information Assistee par Ordinateur"*, pp. 606–623. Elsevier Science Publishers, Amsterdam, NL, 1991.

[14] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. In *Proceedings of National Academic Science*, pp. 7821–7826, 2002.

[15] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 89–97. ACM Press, 1992.

[16] D. Harman. Relevance feedback revisited. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 1–10. ACM Press, 1992.

[17] F. Heylighen. Mining associative meanings from the web: from word disambiguation to the global brain. In *Proceedings of the International Colloquium: Trends in Special Language and Language Technology, R. Temmerman and M. Lutjeharms*, pp. 15–44, 2001.

[18] V.J. Hodge and J. Austin. Hierarchical word clustering – automatic thesaurus generation. *Neurocomputing*, Vol. 48, pp. 819–846, 10 2002.

[19] M. Jarmasz and S. Szpakowicz. Roget's thesaurus and semantic similarity. In *Proceedings of Conference Recnet Advances in NLP*, pp. 212–219, 2003.

[20] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 200–209. Morgan Kaufmann Publishers Inc., 1999.

[21] T. Kamishima. A survey of recent clustering methods for data mining. 人工知能学会誌, Vol. 18, No. 1, pp. 59–65, 2003.

[22] A. Kilgarriff and G. Grefenstette. Web as corpus. In *In Proceedings. of the ACL Workshop on Intellignent Scalable Text Summarization*, 2003.

[23] J. Koenemann and N. Belkin. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 205–212. ACM Press, 1996.

[24] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 170–178. Morgan Kaufmann Publishers Inc., 1997.

[25] A. Lam-Adesina and G.Jones. Applying summarization techniques for term selection in relevance feedback. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 1–9. ACM Press, 2001.

[26] L.S. Larkey and W.B. Croft. Combining classifiers in text categorization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 289–297. ACM Press, 1996.

[27] S. Lawrence. Digital libraries and autonomous citation indexing. Vol. 32, pp. 66–71. IEEE Computer, 1999.

[28] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of ICML-94, 11th International Conference on Machine Learning*, pp. 148–156, 1994.

[29] D.D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 37–50. ACM Press, 1992.

[30] H Li and N. Abe. Word clustering and disambiguation based on co-occurrence data. In *Proceedings of the 17th international conference on Computational linguistics*, pp. 749–755. Association for Computational Linguistics, 1998.

[31] Y.H. Li and A.K. Jain. lassification of text documents. *Computer Journal*, Vol. 41, No. 8, pp. 537–546, 1998.

[32] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pp. 768–774, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[33] Sean McGettrick. Query expansion. Information Sciences and Technology Cource 497, The Pennsylvania State University, University Park, PA.

[34] G. Miller. Wordnet:an on-line lexical database. In *International Booktitle of Lexicography*, 1990.

[35] A.E. Motter, A.P.S. Moura, Y.C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Physical Review E*, Vol. 65, No. 065102, 2002.

[36] M.E.J. Newman. Fast algorithm for detecting community structure in networks. In *Phys. Rev. E 69,2004*, 2004.

[37] S. Oyama, T. Kokubo, T. Ishida, T. Yamada, and Y. Kitamura. Keyword spices: A new method for building domain-specific web search engines. In *IJCAI*, pp. 1457–1466, 2001.

[38] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Manuscript in progress*, 1998.

[39] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, Vol. 435, pp. 814–818, June 2005.

[40] N. Pereira, F. Tishby and L. Lee. Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 183–190, 1993.

[41] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 213–220. ACM Press, 2003.

[42] G. Salton, editor. SMART Retrieval System:Experiments in automatic document processing 書. Englewood Cliff, 1971.

[43] Buckley C. Salton G., Singlhal A. and Mitra M. Automatic text decomposition using text. In *Workshop on Intellegent Scalable Text Summarization*, pp. 10–17, 1997.

[44] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206–213. ACM Press, 1999.

[45] H. Schutze and J.O. Pedersen. A cooccurrence-based thesaurus and two application to information retrieval. In *In Proceedings of RIAO'94 Conf. on Intelligent Text and Image Handling*, pp. 266–274, 1994.

[46] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Survey*, Vol. 34, No. 1, pp. 1–47, 2002.

[47] N. Slonim and N. Tishby. Document clustring using word cluster via the information bottle neck method. In *Research and Development Information Retrieval*, pp. 208–215, 2000.

[48] S. Son, H. Jeong, and J. Dong Noh. Random field ising model and community structure in complex networks, 2005.

[49] D. Susan, J. Thorsten, B. Krishna, and W. Andreas. Sigir 2003 workshop report: implicit measures of user interests and preferences. *SIGIR Forum*, Vol. 37, No. 2, pp. 50–54, 2003.

[50] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP 2004*, pp. 41–48. Association for Computational Linguistics, 2004.

[51] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pp. 491–502, 2001.

[52] M. Wettler and R. Rapp. Computation of word associations based on the co-occurrences of words in large corpora. In *In Proceedings of the 1st Workshop on Very Large Corpora:Academic and Industrial Perspectives*, pp. 84–93, 1993.

[53] D. Widdows and B. Dorow. A graph model for unsupervised lexical acquisition. In *COLING 2002, 19th International Conference on Computational Linguistics*, 2002.

[54] P. Willett. Recent trends in hierarchic document clustering: a critical review. *Information. Processing Manage.*, Vol. 24, No. 5, pp. 577–597, 1988.

[55] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 13–22. Springer-Verlag New York, Inc., 1994.

[56] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420. Morgan Kaufmann Publishers Inc., 1997.

[57] S. Yu, D. Cai, J. Wen, and W. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *WWW*, pp. 11–18, 2003.

[58] 難波 英嗣. 奥村学. 論文の参照情報を考慮したサーベイ論文作成支援システムの開発. 自然言語処理, 第 6 巻, pp. 43–62, 1999.

[59] 佐々木靖弘, 佐藤理史, 宇津呂武仁. ウェブを利用した専門用語集の自動編集. 言語処理学会第 11 回年次大会発表論文集,, pp. 895–898, 2005.

[60] 日本電子化辞書研究所（編）. EDR 電子化辞書 仕様説明書. 日本電子化辞書研究所, 1996.

[61] 神嶌敏弘, 元吉文男. クラスタ例からの学習 : 分類対象集合全体の属性の利用. 情報処理学会, Vol. 40, No. 9, 1999.

[62] 山口翼. 日本語大シソーラス—類語検索大辞典—. 大修館書店, 2003.