氏名（本籍）	山崎久道（東京都）
学位の種類	博士（情報科学）
学位記番号	情第12号
学位授与年月日	平成11年9月9日
学位授与の要件	学位規則第4条第2項該当
最終学歴	昭和44年6月　東京大学経済学部経済学科
学位論文題目	文献情報の蓄積・検索に利用されるファセット分析に基づくシソーラスの開発に関する研究
論文審査委員	（主査） 東北大学教授　福地肇　東北大学教授　阿部四郎 東北大学教授　國分振　東北大学教授　丸岡章 東北大学教授　森杉壽芳

論文内容要旨

第1章　序論

情報は人間や組織の意思決定や行動にとって不可欠の要素であると考えられる。そのためには，情報が円滑に流通するとともに，コミュニケーションが正確に行われることが必要である。現在では、情報に対する人々の接近の仕方は、学問分野に基づくものから、当面の解決すべき問題に基づくものに変化してきている。

こうした状況では、そうした問題を表現する語の存在と意味の確定が重要になってくる。また、文献データベースによる検索は、過去に他者が記述した情報を、語をキーにして現在の自分が探索する、一種の時間・空間を超えたコミュニケーションであると考えられる。それにもかかわらず、現状で主流になっているような自然語を利用した検索方法では、語の意味の変遷、執筆者や検索者の主観による相違、文脈上での意味の差異等がそのまま発現するため、情報の発信者と受信者で正確な概念の一致が実現できない可能性が生ずる。

本研究では、語による検索の正確性の確保を目標としたインデクシング言語を文献データベースに設定することによって、この問題を解決できる可能性があるとの認識のもとに、文献における情報の組織化手法としてファセット分析に基づくシソーラスを開発し、その手順と効果について考察を行った。

第2章　情報管理とシソーラス

本章では、情報管理プロセスの全体と個々の構成要素について述べるとともに、本研究にかかわる重要概念を情報管理プロセスにしたがって定義し、その内容や意義について明らかにした。

情報の利用を効率的に行うには、情報を一定の明確な手順と処理の手法に従って扱う情報管理の考え方が必要である。情報管理プロセスは、収集、整理・加工、蓄積・保管、検索・利用で構成される。このうち、本研究が取り扱うのは、「整理・加工」とそれに関係した範囲での「検索」である。文献の蓄積・検索手法の改善を目的とする本研究では「文献」を、情報の基本単位と見なした。

「整理・加工」プロセスでは、文献で表現されている概念内容（「主題」という）を抽出し（この操作を「主題分析」という）、つぎに主題を表すタグ（ラベル）を付与する（この操作を「インデクシング」と呼ぶ）。語をタグとする場合を「キーワード」、概念にコードを与えてタグとする場合を「分類」と称する。一方、タグの集合でそれがある体系となっているものを、「インデクシング言語」と呼び、インデクシング言語からキーワードを選択してインデクシングする場合を「統制語キーワード」と呼び、そうした統制を行わずもっぱら文献中の語を使用するものを「自然語キーワード」という。インデクシング言語のうち、分類によるものを「分類表」、概念と語の対応づけによるものを「シソーラス」という。さらに、物事を見る観点を導入した情報整理の方法を「ファセット分析」といい、これを分類に適用したものとして「ファセット分類」がある。

一方、「検索」プロセスでは、概念の組み合わせである検索質問を、タグ（キーワードもしくは分類）の集合である「検索式」に変換し、これとデータベース中の文献に付与されたタグとの照合を行う。その結果、当該検索質問で想定された概念の組合せを有する文献を、データベースからどの程度引き出せたかを示す指標を「再現率」と呼び、検索結果に、当該検索質問で想定している概念の組合せを有する文献をどの程度含むかを「適合率」と呼ぶ。再現率と適合率は、ともに検索パフォーマンスの指標を構成する。

第３章　ファセット分析に基づくシソーラス

本章では、従来のシソーラスの形式と事例を説明し、その問題点を検索例を含めて指摘した。つぎに従来のシソーラスの問題点を克服するためには何が必要かを述べ、ファセット分析のシソーラス構築への適用の可能性を提示した。さらに、シソーラスの評価の方法についても具体的に述べた。

従来から作成されているシソーラスは、概念と語の一対一対応は実現しているが、文脈での語の意味の相違を表現できないという大きな問題点を抱えている。そのため、同一のキーワードであっても、インデクシングにおける場合と検索における場合とで、その意味が異なることがある。こうしたことは、求める情報の検索に失敗するという結果を招来する。こうした点を回避するための方策として、シソーラスにファセット分析の考え方を導入するのが有効である。

従来の研究でもファセットの考え方のシソーラスへの適用を論じているが、それらは、用語採集の情報源としてファセット分類を使用する、用語の表示・配列規則としてファセットを利用するなどといったどちらかといったファセット分析の表面的機能をシソーラスの中に実現したに過ぎない。本研究では、まず当該主題分野においてその分野における概念を挙げて厳密なファセット分析を行い、しかる後に用語を採集してシソーラスを構築するという手順を開発した。最後に、シソーラスのそのものの評価を、特に利用者側から適正に行うための基準を作成し、それについて検討した。

第４章　パイロットシソーラスの開発

本章では、ファセットに基づくシソーラスの実際の構築手順を確立する目的で、特定分野を選択し、パイロットシソーラスを構築した。さらに、作成したパイロットシソーラスの性能を評価するためのインデクシングと検索の試験を行った。

主題分野として「料理」を選択して、ファセット分析に基づくシソーラスの構築を行った。料理という分野は、さまざまな観点があり、ある程度成熟した分野であるので、研究上、効果的なファセットの設定が可能である。次いで、メインファセットの設定を行い、用語を採集してシソーラスの形式を整備した。さらに、語の関係、記号法、表示法と構成を決定し、パイロットシソーラスの詳細を作成した。さらに模擬テーマによるインデクシングを行い、その結果に基づいてパイロットシソーラスを修正した。

一方、インターネットのサーチエンジンを利用して、料理分野における特定テーマで検索を行った。この結果をテスト・データベースとして、これに対して、パイロットシソーラスを利用したインデクシングを実施し、再検索して結果を比較した。その結果、パイロットシソーラスを利用した場合には、より高い検索パフォーマンス（再現率、適合率）が得られた。

その理由としては、個々の文献につき、あらかじめパイロットシソーラスに基づいてインデクシングを行ったこと、パイロットシソーラスに基づいて検索式を作成し、そのことにより、検索における照合がパイロットシソーラスを媒介にして行われ、語の表現・表記の多様性が最小限に押さえられたこと、パイロットシソーラスがファセット構造を持っているため、検索式で定義した概念間の関係が、かなり忠実に保存されて、テスト・データベース中の文献と照合されたこと、が考えられる。

ただし、この実験は検索対象となったデータの範囲の選定についてなお課題を残しており、検索対象を明確にして再現率を算定・比較するために、ある規模のデータベース中の文献のすべてにインデクシングを施すとともに、適合文献を事前に目視ですべて抽出しておくといった措置をとった上での検索実験を今後の研究において行うこととしたい。

第5章　結論

本章では、これまでの研究経過と結果に基づき、結論と展望を述べた。

まず、第1に、本研究では、シソーラス構築に関する従来のシソーラスの意義と問題点を指摘した。

従来のシソーラスは、概念と語の一対一対応は実現しているものの、文脈での言葉の意味を保存できない。そのために、同一のキーワードであっても、インデクシングにおける場合と検索における場合とで、その意味が異なることがあり、これがために、しばしば、求める情報の検索に失敗するという問題が発生しうる。

第2に、シソーラスの設計・構築において、ファセット分析の考え方を導入することの効果を明らかにした。つまり、ファセット分析をシソーラスに導入することにより、文脈上の意味を保存した形でインデクシングや検索を行うことが可能になる。

第3に、ファセット分析に基づくシソーラスの実際的な構築手順を開発した。ここでは、シソーラスの作成手順を、従来の用語採集からでなく、概念の整理から始めたことが特徴である。従来のシソーラス作成手順においては、主題分野を決定した後、その分野における基本文献の中から重要と思われる語を次々に収集して、それをグルーピングしてゆくという方法がとられている。しかしながら、このような手法によってシソーラスを構築すると、当該分野の概念の構造が、採集した語のグルーピングの結果により、事後的に形成されることになり、また、シソーラスに収録される語の選択や意味内容がほとんど文献での出現状況にのみ依存することと相まって、当該分野の概念の地図を提示するシソーラスとしては、不完全な形となるおそれがある。

本研究では、まずメインファセットを概念の整理という形で最初に設定し、その中を細分化する段階において、用語を収集してグルーピング、編集を行うという手法をとった。つまり、概念の構造をまず発見し、しかる後に用語の採集に着手するという手順が本研究におけるシソーラス構築の最大の特色である。この手順のほぼ全過程にファセット分析の方法が適用されている。このことにより、当該主題分野における概念の構造が、シソーラスの上で整合性をもって展開されていると考えられ、それとともに、作成したパイロットシソーラスのインデクシング、検索の試験によって、この手順に従うことにより、実用に耐えうるシソーラスを構築できることが明らかになった。

第4に、パイロットシソーラスに、用語と1対1で対応するコードの設定、事項別表示（体系的表示）（Subject Display）と五十音順表示(Alphabetical List）の両方による構成等の形式上の特色を与えた。

第5に、大規模な文献データベースにおいてパイロットシソーラスによるインデクシングを実施し、その有用性評価の可能性と方法論を提示した。

パイロットシソーラスを使わない検索の場合は、集合的概念を直接的に検索式に採用すると、大幅な検索洩れを引き起こすので、個別的概念を、いちいち列挙してＯＲ演算で結合する必要がある。しかも、シソーラスを使わない場合はファセット構造が検索式に反映されておらず、インデクシングもされていないことから、検索結果はさまざまな個別的概念やその異表記が同時に出現するページを探索するだけのものとなり、検索式で指示された概念が検索された文献の主題であることやキーワード間の関係が当初予想されたとおりのものであるかどうかはまったく保証していない。

これに対し、パイロットシソーラスを使った検索の場合は、より高いパフォーマンスを実現できる可能性がある。本研究においては、こうした検索試験の結果の想定と試験方法を提示した。

論文審査の結果の要旨

大量の文献情報をデータベースとして蓄積し、それを効率よく管理・利用するためには、表示概念と適切に対応する索引語を、シソーラスとして体系化することが必要不可欠となる。シソーラスの構築は従来、文献から採集されたキーワードに基づいて行われていたが、検索時に意図する概念を再構成しにくい等の問題点があった。著者は、これに代わる方法として、キーワードの採集に先立ち、過程、実体、空間、時間など、概念分類上の観点を始めに設定する、ファセット分析法をとりあげ、その有用性を示した。本論文はその成果をとりまとめたもので、全編5章から成る。

第1章は序論であり、研究の背景と目的を述べている。

第2章は、収集、整理・加工、蓄積・保管、検索・利用から成る情報管理プロセスについて、その概念と意義を明らかにしている。特に、本研究の中心的部分である整理・加工過程に該当する、主題抽出と索引語付与の手順に関して、自然語キーワードと統制語キーワードの機能を比較、詳述するとともに、統制語による索引付けにおいてシソーラスのはたす役割の重要性を論じている。

第3章では、具体的な事例の検討を通して、従来のシソーラスで生じる、同一の索引語でも索引語付与時と検索時では異なる概念が対応するという問題点を詳細に分析し、文脈を考慮したファセット分析を導入することによりこれを回避できると主張している。さらに、著者は、対象分野において厳密なファセット分析に基づく概念整理をしたうえで用語の収集とその分類・整理を行うことを主張し、その具体的な手順を示している。これは興味深い知見である。

第4章では、対象分野として「料理」を選び、ファセット分析に基づくシソーラスを試作している。ここでは、メインファセットとサブファセットの設定から用語の収集・整理にいたる過程を詳述したうえで、体系表示と50音表示とから成るパイロットシソーラスを作成している。さらに、これを利用した文献の検索実験によって、ファセット型シソーラスの高い有用性を明らかにしている。これらは貴重な成果である。

第5章は結論と今後の展望を述べている。

以上要するに、本論文は、文献情報をデータベース化するためのシソーラスの設計に関して、ファセット分析が有効であることを示すとともに、ファセット分析に基づくシソーラス開発の具体的な手順を与えたもので、情報科学の発展に寄与するところが少なくない。よって、本論文は博士（情報科学）の学位論文として合格と認める。