早稲田大学大学院理工学研究科

博 士 論 文 概 要

論 文 題 目

自動索引作成システムの設計に関する研究

申 請 者
石川 徹也
Tetsuya ISHIKAWA

平成 5 年 5 月

大量の文献・文書等の資料情報を提供する目的のために、書誌データベースの構築・提供がなされている。その際、資料の内容に基づく検索を可能とする索引データとして、索引語と分類番号が作成され、供されている。いづれも書誌データベースの構築時に、索引作成者が資料内容を解読し作成している。この結果、速報性、正確性に欠ける問題点が指摘されている。これに対して1960年代から、索引語作成システムとして資料内容からキーワードを自動的に抽出し、索引作成者に提示、または抽出したキーワードを直接索引語として利用するキーワード抽出システムの設計研究がなされてきた。その方法を大別すると、書誌データベースの構築時に設定する事前抽出・設定方式として、統計解析法、事前設定単語リスト利用法、構文情報利用法があり、検索時抽出方式として、フル・テキスト検索法がある。しかし、従来のどの方式も資料内容の文意を反映した有意味なキーワードの完全な抽出に至らず、不適切な検索結果を生じさせる欠点を持っている。

一方、分類番号を作成するためのシステム化はこれまでに実現していないことから、索引作成者が主題表示データと資料形態特性データを基に分類表を対象にそれぞれの分類番号を検索し、分類番号組み合わせ規則を基に目的とする分類番号を経験的に作成している。この結果、常に正確な分類番号を作成できない状態にある。

そこで本研究において、計算機援用型の効果的な自動索引作成システムを提供するために、上記の２点に対する索引データを自動的に作成する新しい方式を提案した。すなわち、キーワード抽出方式の設計研究のために、文意を反映した有意味なキーワードの抽出を目的に、資料内容を対象に文意解析を行い有意味なキーワードを抽出する方式を、また分類番号生成方式の設計研究のために、主題表示データおよび資料形態特性データの指示の基に、分類表データベースを対象に分類番号の検索を行い、分類番号組み合わせ規則を基に目的とする分類番号を生成する方式の提案とその設計を行った。

研究の対象および研究の方法は、以下の通りである。

キーワード抽出方式の設計においては、日本語記述による科学技術論文の著者抄録を研究の対象とした。研究の方法として、文意を表現する機能の多くは文中に出現する動詞にあることに着目し、動詞の意味概念を基に文意構造を表出する文意解析方式の設計を行い、キーワード抽出指示条件を基に有意味なキーワードを抽出・提示する方式の提案を行い、その設計を図った。そこで、文意解析方式の設計のために、サ変動詞を中心に相当量の動詞を対象に意味概念分析と基本概念記号体系の設定を行い、辞書データの構築を図りこれを有効に利用した。本方式の性能評価のために、主題表示キーワードを対象にキーワード抽出実験を行い、抽出キーワードの再現率および適合率を基に現在実用に供されている事前設定単語リスト利用システムとの比較評価を行った。

分類番号生成方式の設計においては、わが国のほとんどの図書館において資料

配架のために現在利用されている日本十進分類表を研究の対象とした。研究の方法として、日本十進分類表のデータベース化を図り、同分類表の分類番号組み合わせ規則をプロダクション・ルール化し利用する方式を提案し、その設計を図った。本方式の性能評価のために、図書館における実配架資料を対象に分類番号生成実験を行い、実配架資料の分類番号と比較しプロダクション・ルールの完全化を図った。

本論文は７章からなり、各章の概要は以下の通りである。

第１章は序論であり、研究の目的と意義を中心に本研究の位置づけと概要を述べた。

第２章では、従来のシステムの方式と問題点を明らかにした。まずキーワード抽出方式について、統計解析法、事前設定単語リスト利用法、構文情報利用法、フル・テキスト検索法など、従来の方式の検討を行い問題点を考察した。その結果、これまでの方式では抽出されるキーワードは文意に基づかないことから、不適切な検索結果を生じさせる欠点があることを指摘し、文意解析を基に有意味なキーワードを抽出するシステム化が必要になることを明らかにした。また、有意味なキーワードを抽出するために資料内容を対象に形態素解析処理および文意解析処理を行うことの必要性から、日本語文を対象とする従来の形態素解析処理および格文法、結合価文法、概念依存理論法などの構文解析処理方式の検討を行い、問題点を考察し、特に名詞句を欠落することなく文意構造を表出する為の文意解析処理のシステム化が必要になることを明らかにした。分類番号生成方式については、従来分類番号の組み合わせ規則を基に分類番号を自動的に生成するシステムはなく、当システム化の必要性を明らかにした。さらに、これらを組み合わせてはじめて本研究の目的を達成できることを明らかにした。

第３章では、自動索引作成システムの概要と本研究の位置づけを述べた。自動索引作成システムの概要を示し、本研究の位置づけとして、研究の範囲、設計する上での要件を示した。

第４章では、新しいキーワード抽出方式を提案し、次の２つのステップからなるアルゴリズムを明らかにし、これを設計した。ステップ１は文中の名詞句を欠落することなく認識し、しかも文意構造を表出するために動詞の基本概念に基づく文意解析処理の方式である。ステップ２は表出された文意構造を対象に有意味キーワード抽出条件を指示し、目的とする有意味なキーワードを抽出する方式である。そこで、文意解析用辞書データの構築のために、動詞の基本概念体系の設定の必要性から、科学技術論文の抄録を対象に2,147 語の動詞を抽出し、997 語のサ変動詞を対象に意味分析を行い、その基に基本概念記号の体系設定を図った。本方式の性能評価法として、抽出キーワードの再現率、適合率を基に算出し、従来のシステムとの比較を行う方法を示した。

第５章では、新しい分類番号生成方式を提案し、次の２つのステップからなる

アルゴリズムを明らかにし、これを設計した。ステップ1は指示された主題表示データおよび資料特性データを基に、分類名辞を検索する方式である。ステップ2は検索された分類名辞に対応する分類番号を分類番号組み合わせ規則を基に自動的に組み合わせる方式である。このために、日本十進分類表を対象に、分類表の体系、分類番号作成規則の解析を行い、分類表のデータベース化および分類番号作成組み合わせ規則のプロダクション・ルール化を図った。分類表のデータベース化に当たっては、索引作成者が指示するデータに基づき検索処理を行うために、対象とした全ての分類名辞に“読み”を付与した。本方式の性能評価法として、図書館における実配架資料の分類番号に対して全ての機能の確認を行った。

第6章では、第4章、第5章で提案した方式を評価するための実験とその結果を示した。まず、キーワード抽出方式の評価のために、情報処理学会自然言語処理研究会資料の著者抄録の約1年分に相当する33抄録、164文のデータを対象に、その中の動詞317語の意味分析を行い、実験用辞書データの構築を図りキーワード抽出実験を行った。性能評価のために、ＪＩＣＳＴのデータベースに対し実験対象データの検索を行い、そのデータに設定されている索引語を基準キーワードとし、本方式による抽出キーワードと現在実用に供されているシステムによる抽出キーワードについて、再現率および適合率を基に両システムの性能検定を行った。この結果、本方式が従来のシステムに比べ有意味なキーワードを抽出できることの確証を得た。次いで、分類番号生成方式の性能評価のために、図書館情報大学付属図書館に配架されている24冊の図書資料を基に、主題表示データと資料形態特性データを指示し分類番号生成実験を行った。この結果、生成分類番号を同資料の分類番号と比較し、極めて正確な分類番号を生成することの確認を行った。

第7章は結論であり、本研究において提案した方式が、索引語作成、分類番号作成を行う上で、従来のシステムに比べ充分有効に機能することの結論を示し、同時に残された今後の研究課題を示した。