【209】

氏名(本籍)	荒木啓介(埼玉県)
学位の種類	博士(工学)
学位記番号	博乙第783号
学位授与年月日	平成4年3月25日
学位授与の要件	学位規則第5条第2項該当
審査研究科	工学研究科
学位論文題目	化合物の名称解析と立体構造処理システムの研究

主査	筑波大学教授	博士	電子情報工学系	藤原　譲
副査	筑波大学教授	博士	物質工学系	内島俊雄
副査	筑波大学教授	博士	電子情報工学系	鈴木　功
副査	筑波大学助教授	博士	電子情報工学系	大保信夫
副査	筑波大学助教授	博士	電子情報工学系	北川博之

論文の要旨

(1) 化合物名称解析システム

物質に関する学問・技術である化学は，過去の情報や知識の上に立って新しい知見を加えていくという色彩が濃いが，そのためには，化学物質を正確に同定し，名称や構造上の様々な特徴から検索可能なデータベースとして蓄積しておく必要がある。そのためには，分子構造を標準化された原子結合表の形で表現することが前提となる。

化合物は，その分子構造が一種の人工言語である体系名によって表記されるので，計算言語学の手法を用い，計算機処理可能とした辞書（体系名要素の名称等とその構造）とプログラム化した文法（命名規則）により解析し，分子全体の構造を組み立てるシステムの作成を試みた。

システム分析に当たっては，様々な化合物の体系名について，その構成要素や命名規則を詳細に調査分析し，体系名を解析し組み立てるアルゴリズムを確立した。

次いで解析辞書のコーディング，システム設計，プログラムの作成，テストと，5年を要して全体システムを開発した。

また，米国CASのRegistry Systemに比べて20年も遅れた開発であるが，CASでは欠けている化学構造の立体化学的な表現を，より完全なものとする方式を思案し実装した。立体化学の記述情報を持った体系名を100%近い精度でその原子結合表に変換し，グラフ理論的な手法を基礎にしたプログラムにより標準化して，次に入力される構造と照合して自動的に同定，追加するシステムとして作り上げた。

基本的な方法は，立体化学を記述するための世界的な標準とされている，Cahn, Ingold, Prelogの優先順位規則を計算機プログラムとし，入力された化合物体系名中の立体記述子の意味を解析して，該当する不整原子の立体配置を，立体特異的な原子結合表に書き込むというものである。

不整原子に結合した配位原子の優先順位を決めるための手続きの忠実なアルゴリズム化と，それ自体が立体化学的な意味を持つ原子結合表の思案および正しく結合表に書き込むための手続きのプログラム化に今までにはない新規性がある。

化学物質のデータベースとして共に必須な体系名と分子構造を別々に入力する工程を省力化し，JICST程度の規模の，しかも全科学技術分野を対象とする総合センターにおいても，文字入力のみによって，化学に特有な構造情報を扱うことが実用的に可能であることを実証した。

基本的な研究は1976年頃から始めたが，本格的な研究およびシステム分析，システム開発は，1981年9月に開始された科学技術庁の振興調整費プロジェクト「ネットワーク共用による化合物情報等の利用高度化に関する研究」にこの構想が採用され，体制が整えられてから具体的に進展し，解析辞書35,000項目以上，プログラム規模150Kステップのシステムとして完成した。

ファクトデータベースネットワークの中心的な化合物辞書としてその対象化合物を登録し，別途開発された専用の検索システムを通して各専門データベース間の渡り検索を可能とすると共に，1990年からはJICSTの文献データベース作成システムと連結し，物質索引のオーソリティ辞書としての役割を果たしている。

1991年10月時点で約380,000化合物を蓄積し，JOIS, JOIS-Fを通して一般にサービスされている。

この，言語処理の手法を利用した化合物システムの方法論的な基礎ともなり，経験を積むうえでも有用であった。

(2) 漢字ーカナ変換システム

特殊法人日本科学技術情報センター（JICST）では，1969年から漢字テレタイプライターによる科学技術文献抄録の入力，科学技術文献速報の機械編集を行ってきた。

また1972年からは，編集した磁気テープを一般に提供し，企業等における文献情報の機械検索に供すると共に，1975年からはJOISオンラインサービスも開始した。

しかし，検索用磁気テープもオンラインサービスも，当時のハード環境や計算機の処理能力条件等から，データは全て英字，カナ文字モードであり2バイト系の科学技術文献データベースのマスターファイルから，1バイト系に変換して用いられてきた。

この際，文献抄録の書誌事項（雑誌ID，巻，号，頁，年等）は自動的に1バイト系になり，雑誌名も資料マスターのカナ名称から転送可能であり，キーワードも，シソーラスのフリガナから自動的に与えることができたが，論文のタイトル，抄録文の1バイト系への変換は困難であった。

そのため，抄録は欠落することとなり，論文タイトルは原文が欧文のものはそのまま1バイト系にして移し，ロシア語は翻字した形のタイトル文とせざるを得なかった。

原文が日本語のみのものは別途入手によりカナ表題を作り，パンチ入力した。これはデータベースを作る側にもユーザーにとっても不便を強いるものであった。筆者はこれを，日本語文の自動カ

ナフリ技術を開発する事により解決する事を考え，1976年から調査，検討を開始，1978年秋には，コンパクト，高精度なプログラムシステムとして完成，1979年以降実用化され，上記の不便を解消するとともに，抄録索引を担当する知的な専門家による単調なカナフリ作業を無くし，作業能率の向上，経費節減およびサービス製品の改善の双方で効果を上げている。

(3) 日本語文からのキーワード自動抽出

先に開発した漢字ーカナ変換システムにより，漢字かな混じりの日本語文は読むのに支障が無い程度に分かち書きし，カナ変換される。このままでも約86%程度の切り出された文字列はそのままキーワードとして使用できる。しかし分かち書きは，接続詞，助詞，助動詞，動詞，形容詞等のいわゆる用言などによってのみ切断しているため，漢字のみから成る文字列は切断されないまま残り，キーワードとしては不適切なものを多く含む。

日本語文字列中の接辞（上，中，下，各等）に注目し，また名詞が並んだ文字列のパターンを分析した。前者については，接辞文字が，接辞としてでない用語を作るケースを収集して切断をパスさせることとした。

上については，上陸，上昇，向上　等

後者については，連続する各名詞列の中で，数が少なく網羅できると思われる語を収集し，切断キーとすることとした。

ラット	肝　臓	ミトコンドリア	ATPアーゼ
動物名	臓器名	細胞構成体名	酸素名
（多数）	（有限）	（有限）	（多名）

臓器や細胞構成体名で切ることにより，全体がバラバラになる訳である。

約10万の科学技術論文タイトルからこのような接辞及びその接辞にならない熟語を集め，また体言のみの文字列のパターンを調べてその切断方針を検討した。

約7,000語の切断辞書を収集し，効率の良いプログラムを作成して，精度99.3%以上のシステムとして完成した。

JICST科学技術文献データベース1975年ファイルから1991年ファイルまで，延べ550万件の論文タイトル及び抄録文について適用し，2,000万語以上のフリーキーワードを抽出，JOISによる検索に供されている。

この中には，最先端の学術用語が含まれ，ユーザーになじみの深い自由なキーワードによる的確な検索を可能としている。

審　査　の　要　旨

化学の研究開発に必要な情報の中心となる化合物名称と化学構造の入力，相互変換の方式を確立し，それに基づく化合物入力システムを開発した。主たる要点は次の通りである。

1．化合物の命名法を国際標準であるIUPAC命名法に準拠し，その未整備，不正確な部分を含めて

計算機処理用の体系的命名法を確立した。

2．体系的名称または慣用名から化学構造を生成し，結合表の形で表現する方法を確立した。

3．主体化学を自動判定し，体系的に表現する方法を確立した。

4．以上のため名称解析と部分構造生成のため40,000項目を超える辞書を作成した。

5．以上に基づき，化合物名称及び化学構造のデータベースを構築管理するシステムを設け開発した。

6．389,000件以上の化合物辞書を構築し，実用に提供している。

7．化合物3次元構造予測システムを始め，化学の研究開発に必要な応用システムを開発しつつある。

8．関連して漢字ーカナ変換システムおよび日本語文からのキーワード自動抽出システムを開発し，化合物DBシステムのみならず広く実用に供されている。

よって，著者は博士（医学）の学位を受けるに十分な資格を有するものと認める。