文献検索システム NKWIC におけるキーワードの自動抽出について

名大・大型計算機センター　○飯田三郎
名大・土木　正員　梶田建夫
中部工大・計算機センター　水島章次
名大・土木　正員　成岡昌夫

1. はじめに

文献検索システム NKWIC は，データバンクとして有限要素法に関する文献情報，機械振動に関する文献情報を持つ検索システムで，本年5月より名大大型計算機センターで検索サービスが行われているものである[1]。前報告[2]において，このシステムの概要とその利用について述べたが，そののち実際にサービスするにあたり，システムの改良をいくらが行った。有限要素法に関するデータには自然語のキーワードがつけられているが，機械振動に関するデータには分類コードしかなく，自然語で検索することができなかった。このため，個々の文献の表題から機能語，および，不要語を除き，さらに複数形の単語の単数化などの処理を施してキーワードを作り出す自動索引の機能を開発した。データバンクの作成において，情報の蓄積を継続して行うことは重要なことであり，このためにはデータの作成を容易にできるようにする必要があるが，この点に関しても，この自動索引の機能は有効であり，一文献のデータは，題名，著者名，書誌事項のみでよいこととなる。しかし，表題中の単語をキーワードとすることは検索の効率の点では問題があるので，各キーワードに関連するキーワードのテーブルを作り，検索においてそれを利用できるようにした。

また，良質なデータベースを作成するためには，カードにパンチされた原データの誤りをできるだけ取り除く必要がある。このためのエラーチェックの機能についても開発を行った。

2. 基本語彙

表題を単語に分解し，それぞれの単語すべてをキーワードとして扱うこともできるが，このようにすると，キーワードの転置ファイルが非常に大きなものとなり実用的ではない。したがって，まず表題より抽出された単語のうち，接続詞，冠詞など明らかにキーワードとなりえないものは除いた。また，接尾辞の違いだけで別のキーワードとして扱われることとなるので，これも一つのキーワードにまとめることにした。

機械振動に関する文献約17000件の表題中には，約182000語の単語があり，その種類は約14000語であった。この14000語を接尾辞による派生形としてまとめると，その総数は約3500語となった。これを基本語彙とし，それぞれの基本語彙と接尾辞との関係は別に表としてもった。

つぎに基本語彙の一部を示す。

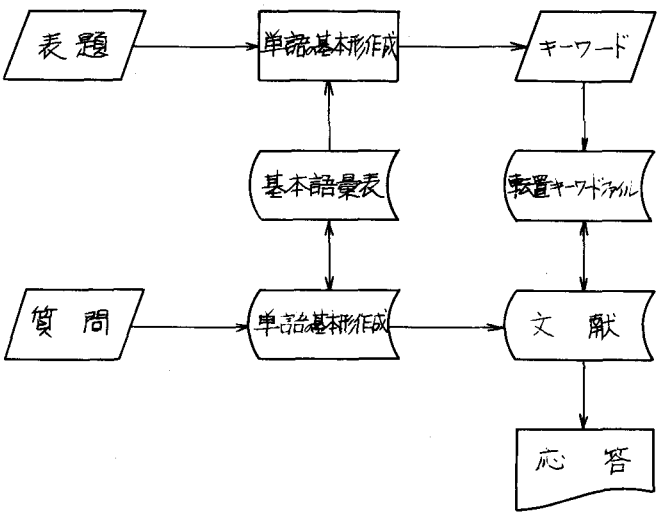
```
MASK                    * ED ING S
MASONRY                 *
MASS                    * IVE Y IF ES
MAST                    * S
MASTER                  * S ED ING FUL LY Y
MAT                     * S TED TING
MATCH                   * ES ED ING
MATERIAL                * S LY ITY IZE IZED IZING IZATION ISE ISED ISING ISATION
MATHEMATIC              S AL IAN
MATHIEU                 *
MATRICES(MATRIX)        *
MATRIX                  * ES
MATTER(MATERIAL)        * S ED ING
MAXIM                   A AL UM UMS ALITY IZE IZED IZING IZATION
                        ISE ISED ISING ISATION
MAX(MAXIMUM)            *
MAXWELL                 * S IAN
MEAN                    * S ER EST T ING INGLY LY
```

基本形と接尾辞とは，どれを接続しても同じ概念となるもののみをまとめており，接尾辞により概念の異なるもの，たとえば，lessで始まる接尾辞をもった単語などは，lessまでを基本形としそれ以降は接尾辞表によることにした．この基本語彙表は，キーワードの自動抽出に使用できるほか，基本形と接尾辞とを接続することにより，表題に限れば原データの誤りの検出にも使用できる．

なお，上述したような基本語彙は，機械振動の分野の文献から作成したが，有限要素法の文献約10000件に関するキーワード約4000種類を全て含んでおり，両方のデータベースに対して使用できるものと思われる．

3．自動索引および検索

表題よりキーワードを抽出するためには，前に示した基本語彙をファイルに入れておく必要がある．文献データの表題を読み，これからストップワードを除去し，単語の連結体を抽出する．[3,4] つぎに，基本語彙を用いてこれを基本形に変換し，キーワードとして転置ファイルに格納しておく．したがって，質問文（この場合は単語のみでなく一般の文章でもよい）も入力と同時に基本形に変換され，それが単語の連結体であれば，それらの単語の論理積を意味する検索が行われる．



この状態を図示したものが右の図である．

表題中の単語をキーワードとして検索することは，表題により論文の内容がすべて表わされているわけではないので，十分な検索を行えないことは明らかである．また，一般に用いられているようなシソーラスにより，検索の効率をあげることはできるが，このシソーラスを作り出すことはかなり困難なことである．

我々のシステムにおいては，個々の基本語について，同意語，関連語の表を作り，これを基本語彙と同じファイルに入れておき，検索のときに用いられるようにした．同意語，関連語の表を作ることは，シソーラスの作成と同様にいろいろ問題があるが，ここで決められた同意語，関連語は検索の途中において表示されるのみで，他のシソーラスを用いたシステムのように，これで検索を行うことはしない．すなわち，あるキーワードについて検索を行い，このキーワードに関する同意語，関連語が知りたければ，コマンドでこれを要求すれば出力される．したがって，質問者はこの出力されたキーワードを見て，必要と思われるものがあれば，つぎにそのキーワードについて検索の要求を行い，必要がなければつぎに移るようにすればよい．

つぎに基本形とそれに関する同意語，関連語の例を示す．

```
PLATE
    see      DISK,SLAB,FLOOR,GRID
QUADRILATERAL
    see      PARALELOGRAM,RECTANGULAR,RHOMB,SQUARE
STABLE
    see      EQUILIBRATE,STEADY
VAULT
    see      DOME,SHELL,SPHERICAL
```

4. システムの維持

はじめにも述べたように，我々のシステムはすでに名大大型計算機センターの利用者に公開されており，このために長期間にわたるシステムの維持を行わなければならない．今後もデータの作成および更新が行われていくわけであるが，これが継続して行われていくためには，前に示した自動索引を含め，種々の機能がシステムに付加されなければならない．その一つとして原データのパンチミスを検出する機能は，自動索引に用いられた基本語彙表により行うことができる．

原データのパンチミスを検出する方法としては，同じ誤りを2回繰り返すことはまれであるという1回ヒット法（once hit method）[5]を使用することが考えられるが，我々のシステムでは基本語彙表内の単語とのマッチングによる方法[6]を採用した．

入力された原データを単語に分解し，その単語と基本語彙表内の単語とのマッチングをとる．もし完全にマッチングがとれないときには，その警告と共にもっとも似かよった単語を出力する．このとき，入力した単語がシステムの出力した単語の誤った形のものであれば，オペレータ（研究者）の指示のもとで自動的にその誤りを訂正する．また，入力した単語が正しいものであれば，その単語は基本語彙表にないものであるから，別に接尾辞および関連語をしらべて基本語彙表を更新する．上に述べた似かよった単語とは，構成文字およびその並びがもっとも近い単語をいい，それを決定するアルゴリズムはつぎのようである．

ステップ 1　入力した単語を構成する文字の組にもっとも近い基本語を選ぶ．これは複数個でもよい．

ステップ 2　ステップ1で選び出された単語のうち，文字の並びが最も近いもの一つを決定する．
このステップにおいて，文字の並びが同じときには正しい単語である．

この方法を既存のデータバンクのデータに適用したところ，once hit methodでは検出できない誤り，すなわち，同じ個所で同じ誤りをした単語でも検出でき，良好な実験結果を得ている．

5. おわりに

以上，文献検索システムNKWICにおける自動索引と誤り検出のシステムについて述べたが，これらはまだ十分なものではなく，利用を続けながら今後も改良を行っていくものである．今日の文献情報の増加は急激であり，我々はこのような検索システムに頼らざるを得ない状態となっているが，そのような要求に対応できる検索システムとするために，今後も研究開発を続けていかなければならないと思っている．

最後に，同意語，関連語の表の作成に協力頂いた名古屋大学工学部土木工学科助手 馬場俊介氏，有住康則氏および同大学院生 清水茂君に感謝の意を表わします．

参考文献

1) 飯田，梶田，水島，成岡：文献検索システムNKWICについて，名大大型計算機センターニュース，Vol. 8，NO. 2，PP. 131～142

2) 飯田，梶田，水島，成岡：有限要素法および機械振動に関する文献検索システムNKWICについて，電算機利用に関するシンポジウム講演概要集，土木学会電算機利用委員会，1976年11月，PP. 76-79

3) 有川：MIR-RF情報検索システム，情報処理学会全国大会予稿集，昭50，NO. 42

4) A.E. PETRARCA，W.M. LAY: The Double-KWIC Coordinate Index，Journal of Chemical Documantation，Vol. 9，NO. 2，PP. 256-260

5) 山田，他5名：コンピュータリーダブルな化学文献データの誤りチェック，トクメンケンキュウ，1975年3月，PP. 95-99

6) F. J. Damerau：A Technique for Computer Detection and Correction of Spelling Errors，CACM，Vol. 7，No. 3，PP. 171-176