D-36　検索システムにおける仮名文字の自動変換システム

The automatic conversion system of Japanese-syllabary Characters in a search engine

根本 良明†
Yoshiaki　Nemoto

1．はじめに

近年，インターネットの利用人口は急増し、インターネットへアクセスする人は研究者等の一部の人間から専門の知識を持たない一般人の方が多くなっていている。その一般人のほとんどは，インターネットを楽しむためのホームページを探すために検索サイトを利用している。

本研究では，利用者の中でもコンピュータの操作に慣れていない人達でも，快適にインターネットを利用するために，仮名文字を自動変換する検索システムを検討し実装を試みた。

2．検索システム

2.1　一般的な検索システム

インターネットに接続し，web 上を見て回る場合に検索システムは，とても有用なシステムである。goo や infoseek，Google，Lycos，Yahoo! Japan に代表される検索サイトは膨大なインターネット上のホームページの情報を所有しており，キーワードを与えるだけで，利用者が求めるホームページを探し出してくれる。しかし，キーワードと同じ文字しか検索しないため，平仮名と漢字で検索した場合には，検索結果に大きな差が出てしまう。 例えば，“検索システム”という文字をキーワードとして与え Google で検索してみる。すると，133,000 件という膨大な数が検索結果として提示される。次に“けんさくしすてむ”という文字で検索してみる。結果は該当無しという文字が表示されるだけである。他の検索サイトで同じことをした結果は，“検索システム”よりも“けんさくしすてむ”の方が該当数は極端に少ないということである(表１)。このことにより，漢字と仮名という違いだけで同じ意味を持つ言葉でも結果が変化するということがわかる。これは，漢字と平仮名，片仮名では文字自体の情報が異なるからである。では，“検索システム”と“けんさくしすてむ”ではどのように情報が違うのか表にて比較してみる。

表２からわかるように，漢字，平仮名，片仮名では文字コードが異なっている。そのため，“検索システム”と“けんさくしすてむ”では検索結果が異なるのである。また，文字コードには SHIFT-JIS コード以外に，JIS コード，EUC コードが存在し，これらのコードにおいての文字の情報は全て異なっている 4)。

2.2　子供向けの検索システム

子供向けの検索システムの代表として、Yahoo!きっずが存在する。この検索サイトは、対象を小中学生に絞り、漢字には括弧で読み仮名を表示している。そのため漢字を読めない子供向けとしてはいいかもしれないが、その他の検索システムの部分では、Yahoo! Japan と変わらないため、検索結果の内容に差が出てしまう。

例えば、”けんさく”という単語で検索を実行すると、検索結果において”けんさく”の単語しか検索しておらず、しかも検索結果数は 11 件である。これでは、漢字に読み仮名が表示されているだけで、Yahoo! Japan よりも検索能力において劣っている。

表１：検索結果一覧

サイト名	検索文字	
	検索システム	けんさくしすてむ
Google	133,000	0
infoseek	24,456	926
Lycos	235	68
goo	93,578	1
Yahoo!Japan	166	79

表２：文字コード一覧(SHIFT-JIS)

検	8C9F	け	82AF
索	8DF5	ん	82F1
シ	8356	さ	82B3
ス	8358	く	82AD
テ	8355	し	82B5
ム	8380	す	82B7
		て	82C4
		む	82DE

3．自動変換システム

3.1　自動変換するにあたって

実際に，仮名入力から自動的に漢字へ変換するには大きな壁が存在する。それは，同音ではあるが文字の異なる漢字が存在するということ，一般的には変換することが不可能な各ユーザーが使用している特種文字などである。

同音異字はユーザーが任意に変換する場合はユーザーが選択すればいいのだが，自動的に変換してしまうと，目的の単語とは違う単語に変換してしまう可能性

†茨城工業高等専門学校
Ibaraki National College of Technology

がある。

特種文字とは，ユーザーが任意に決めた読み方で変換されるように登録された単語のことである。これにより，一般的には変換されないような仮名入力でも変換が可能となる。

3.2 同音異字の解決方法

この場合の解決方法の一つとしてとして，段階的検索が挙げられる。この方法は図１に示すように、一度目の検索で該当する漢字の単語組み合わせ一覧を表示し，その中に該当する単語があった場合は，それをクリックすることにより検索結果を表示するという方法である。これによって，一度に検索してしまうと同音異字の単語も含めて検索してしまうということがなくなり，ユーザーが必要としている情報を絞り込んで検索することが可能となる。しかし，段階的に検索したり、複数のキーワードが与えられた場合に検索時間が長くなるということ，検索システム側に登録されていない漢字の組み合わせの単語は検索されないという欠点がある。

もう一つの方法として，使用頻度の差から順位をつけ検索する方法がある。これは，web サイト上で使用されている単語の中から，使用頻度の多い単語を上位の順位に位置付けすることによって，検索結果を上位のものから優先的に表示する方法である。しかし，頻繁に使用されない単語を検索する場合には、優先順位が低いため、検索結果数によっては検索結果の中から探し出すことすら困難になってしまうという欠点が存在する。

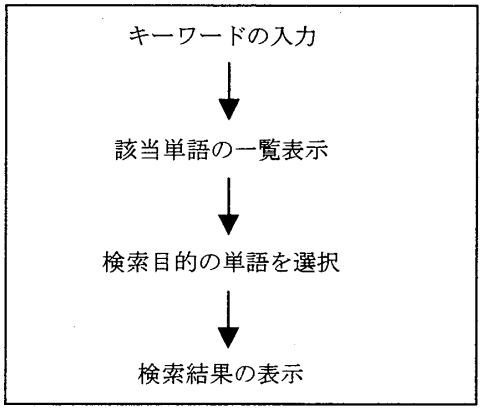

図１．検索の流れ

3.3 特種文字の解決方法

特種文字は，文字コードの空いている部分にユーザーが登録することによって使用可能となる文字なので，各ユーザーごとに異なる文字が登録されていたり，何も登録されていないということがある。したがって，基本的には使用しているユーザー以外には変換が不可能である。しかし、ほとんどの特殊文字はわずかなユーザーにしか知られていないため、検索に使用されることは滅多にないので無視するとにしてもかまわないと考えられる。

4. 単語追加機能

登録されていない単語検索や特殊文字の検索を実行できるようにするために、システムにユーザー側から単語登録を出来る機能を組み込む。

登録内容としては、登録する単語、その単語の読み方である。これにより、同音異字での一段階目での該当単語表示に漏れてしまう単語や、特殊文字などが検索することが可能となり、幅広いユーザーが利用できることになる。しかし、ユーザーなら誰でも登録可能であるので、心許無いユーザーにより、悪戯が目的での無意味な登録により、検索時間の長時間化が考えられる。

そこで、図２のように e-mail のアドレス入力することにより、その e-mail アドレスへ ID を発行し、ID を入力したユーザーにのみ登録権限を与えるようにする。これによって、多少は悪戯目的の登録が解消できるかもしれない。

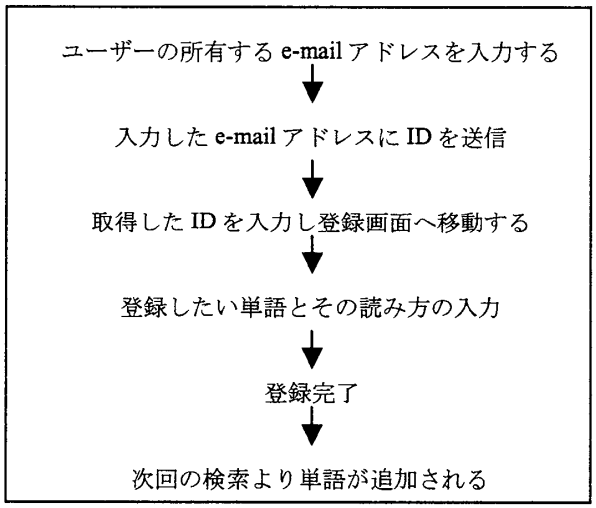

図２．新規単語登録の流れ

5. 実装

自動変換システムの部分は Perl 言語により実装した。検索システムの主要部分は自動変換システムを調整するためにフリーの検索システムを利用した。

6. まとめ

自動変換システムにおいて変換の解決方法を述べた。今後は，このシステムの部分を中心に研究を進めたいと思う。また，仮名入力だけではなく，英字入力からの自動変換にも対応するようにできればさらに利用しやすくなるだろう。

参考文献

1) http://www.goo.ne.jp/
2) http://www.google.co.jp/
3)http://www.yahoo.co.jp/
4)http://ash.or.jp/code/codetbl2.htm