# Document Query by Example 文書の例示検索

5P－7

湯浅聖記　堀口雅祥　木村浩明　芝野耕司
東京国際大学　商学部　経営情報学科

## 1　はじめに

インターネットの爆発的な普及によって、これまで情報検索に縁のなかった一般の人が情報検索を行いたいという要求が生まれている。現在の代表的な情報検索システムは、キーワード検索と全文検索が存在するが、どちらの検索法でも非専門家の検索要求に応えることは難しい。

現在のキーワード検索においては検索に用いるキーワードの選定には、サーチャー的な専門性が要求される。全文検索においては、より一般的ではあるが、適切な検索要求を指示することは依然として、一定程度の専門性を要求されることには変わりない。そこで専門性を必要としない検索システムが求められる。

全文検索システムの代表的なアクセス法としては、シグナチャの研究は存在するが、日本語文書を対象として実証的に構成した例は少ない[1]。同時に曖昧な検索を許容する例も少ない。この論文では、検索対象の例示となり得る文書を指示し、この文書に類似した文書をその類似度のランク付けを含めて検索可能な検索・アクセス手法を提案する。

この論文では、１文字漢字と２文字漢字の度数分布を統計的に求め、基本漢字シグナチャを階層的に構成する。各文書毎に１文字と２文字の基本シグナチャ上の度数分布を作成し、特徴ファイルとする。類似度検索は指定文書と検索対象文書との距離を特徴ファイルを用いた情報量基準をもとに計算し、この値で行う。

## 2　日本語シグナチャの構築

### 2.1 日本語シグナチャ構成の問題

全文検索を高速化するための手法としては、シグナチャ法がある。シグナチャの構成法は、文字を単位とする構成法と単語を単位とする構成法とがある。これらは、文中にある文字や単語を0と1とからなるビット列で表現し、これを利用し検索する方法である[2]。このシグナチャの構成法は、欧米語の次の性質に依存している。すなわち、欧米語では文字種が少ないこと及び単語は複数の文字からなり、単語間は空白又は約物によって区切られ、簡単に単語抽出ができるという性質を基本としている。日本語文書の場合、文字種の多さ及び単語抽出の困難さから、シグナチャ法の適用には、日本語文書の性質を考慮する必要がある。

一方、漢字は表語文字（logograph）である。すなわち、一つの漢字が基本的には一つの単語を表す。この性質を活かす必要がある。この点では、文字単位と単語単位のシグナチャは、漢字の場合は基本的には同じものとなる。しかし、実際の日本語文書では、熟語としての利用が多いことを考えると、日本語文書での単語シグナチャとしては、1文字単位のシグナチャに加えて、2文字の熟語の多さを考慮すると、2文字単位でのシグナチャも考慮する必要がある。

日本語を含む漢字文化圏で用いられる文字種は、欧米に比べて極めて多い。例えば、ISO／IEC 646

---

Document Query by Example.
Masaki Yuasa,Masayoshi Horiguchi,Hiroaki Kimura,
Kohji Shibano
Tokyo International University

IRV (ASCII) では、7ビット又は8ビット1バイトを用い、94字からなる文字コードを基本とするが、日本の情報処理では、94×94=8836の7ビット又は8ビットバイト2バイトを用いる必要がある。実際にJIS X 0208で規定される漢字コードでは、6879文字に符号を与えている。ASCIIをベースとした文字シグナチャが12オクテットになるのに対して、JIS漢字をもととした漢字シグナチャは1000オクテット以上となる。そして2文字シグナチャは、1Mオクテットとなる。このように単純にシグナチャを構成すると、シグナチャインデックスのサイズが百倍近くなるため、効率的ではなくなる。

こうしたことから、日本語シグナチャの構成にあたっては、何らかの形で、シグナチャを圧縮することが必要となる。

### 2.2 漢字出現頻度調査

日本語文書中で用いられる漢字は、現代文においては、常用漢字表に含まれる1945文字を基本とし、固有名詞などで用いられる漢字がこれに加わり、3000文字もあれば、90%以上の文字を含むこととなる。一方、康　字典、諸橋大漢和などの漢字字書には、50,000文字以上の漢字が収録されている。JIS漢字コードには、地名漢字を多く含み、それぞれの漢字の出現頻度には大きな偏りがある。例えば、"岾"は京都市左京区浄土寺の地名としてだけ1987年まで用いられていた。常用漢字表においても、日本国憲法など一部の法律文書でだけ用いられる漢字を含んでいる。

## 3 例示検索

### 3.1 検索

検索は、まず最初に例示として与える文書を指示し、指示された文書に類似した文書を検索することによって行う。この検索は、データベース中より同じ基本シグナチャパターンをもつ文書を検索することによって行う。この検索によって得られる集合を与えられた文書に類似した文書とする。

### 3.2 類似性とランク付け

このようにして得られた類似文書に対して、類似度を計算する。この類似度の計算には、各文書の特徴ファイルを利用し、情報量規準をもとに、文書間の距離を算出することによって行う。

この情報量規準をもとにした文書間の距離の算出は、各シグナチャビットの出現頻度の絶対差の合計を用いる。

## 4. おわりに

この例示検索は、シグナチャを利用することで、記憶容量が減り、検索スピードを上げることができる。その上、情報量規準を利用することで、文書の類似性を表現できる。

## 5. 参考文献

[1] 玉置　志津他:シグネチャ法を用いた日本語文書検索システム,平成7年前期情報処理学会全国大会,第50回,4−39,4F−1

[2] 小川隆一他:フルテキスト・データベースの技術動向,情報処理,Vol.33,No.4,pp.404-412(1992)