多数の語を用いた検索質問の作成と評価

○石田 栄美
宇陀 則彦
石塚 英弘
根岸 正光
山本 毅雄

Formulation and Evaluation of Queries Using Many Words

○Emi Ishida
Norihiko Uda
Hidehiro Ishizuka
Masamitsu Negishi
Takeo Yamamoto

The present study describes formulation and evaluation of queries for large objects such as entire fields of study, for Japanese textual data.

The proposed method consists of decomposing Japanese textual data in several content groups and taking any word which shows markedly different relative frequency for a pair of groups as specifying the group in which it appears. The set of words is then used for retrieving texts belonging to the group.

Abstract texts, in Japanese, each of about 2000 conference papers in three fields of study (information processing, agricultural chemistry and civil engineering) were downloaded from the GAKKAI database at National Center for Science Information System. About 1000 each of them were used for formulating queries, and the rest were used for evaluating the queries. In formulating the queries, texts were first decomposed into words by using MHSA (a multiple-hash screening algorithm for decomposing texts into words when there is no obvious delimiter). Relative frequencies of words in all the groups were measured and analysed. Two word lists, a general word list which includes all words appearing in each group, and a specific word list which lists words in a group whose relative frequency is markedly high or low in at least one group, were formed for each group.

By using the specific word lists for retrieval, it was found that evaluation set of texts of the above three fields can be classified to the original group with an accuracy of about 96%.

1　はじめに

文献検索を行う場合、比較的小さな主題について検索する場合が多いが、往々にして、「情報処理」、「論理学」など大きな分野全体についての検索を行いたいことがある。小さな主題の検索の場合は、数個ないし数十個の比較的少数の検索語で十分検索できるが、大きな分野の検索の場合には、その分野を網羅する多数の検索語を手作業でリストアップし、これを評価するのは困難である。

本研究では、大きな分野を検索するための検索質問の自動的作成手法を確立することを目的とする。

自然言語の語彙解析による検索、自動分類などは、Salton[1] の研究その他があるが、これらは主として文献の小分野への分類を目標とするものである。本研究ではこれに対し、大きな分野の間の分類、判定を日本語情報に対して行っている点が特徴である。

本研究では、情報処理、農芸化学、土木という異なる３つの分野の文献を対象とし、これから各分野の検索用単語リスト各２種を自動的に生成する。これらを用いて同じ分野の別の文献を検索し、単語リストの作成手法の有効性を検討する。自動的に作成された多数の語を分野検索用の単語リストと呼ぶ。

単語リストは、検索したい分野の大量のテキスト（抄録）から自動的に作成する。単語の切り出しは、複数ハッシュふるい分け日本語わかち書きシステム MHSA(Multiple Hash Screening Algorithm)[2][3]

によって行う。MHSA は、文を構成するあらゆる語の組み合わせから辞書にある語を全て切り出すため、フルテキストからキーワードを網羅的に抽出できる。

大きな分野はその中に小分野を複数含んでいるため、特定の小分野にかたよらない網羅的な単語リストが必要である。

単語リストは、MHSA によって切りだした単語をすべて含んだ「分野総単語リスト」と分野を特定しない単語を省いた「分野特定単語リスト」の２種類を作成する。分野を特定する単語かそうでないかは、出現率の差異によって判断した。

本研究では、学術情報センターの学会発表データベースのうち、情報処理学会、日本農芸化学会、土木学会の３分野の各々約 1000 件の文献を用いて単語リストを作成し、同じ３分野の別の約 1000 件に対して検索実験を行った。

2　単語リストの作成方法

2.1　単語リストの概要

単語リストは、MHSA によって切りだした単語をすべて含んだ「分野総単語リスト」と分野を特定しない単語を省いた「分野特定単語リスト」の２種類を作成する。分野特定単語リストは、３つの分野での出現率に大きな差のある単語を集めたものである。１つの分野で特に出現率が高い単語だけではなく、いくつかの分野で共通に（専門語として）使われている単語も採用されるようにした。

2.2　分野総単語リストの作成

分野総単語リストは、文献中の単語を MHSA によって全て切りだしたものである。以下情報処理分野の総単語リストに頻度をつけたものを示す。

```
11153                の
8142                 る
7201                 を
6987                 に
4546                 と
4357                 て
        （中略）
288          アルゴリズム
287           プロセッサ
287                 また
283                   場
264                   率
259          プログラム
259                 手法
258                 実行
257                 解析
256                 には
254                 報告
254                 時間
216          ネットワーク
        （中略）
10      データベースシステム
10      スーパーコンピュータ
10      コンピュータシステム
10        ユニフィケーション
10            プロセス間通信
10              リフレッシュ
10              ボトムアップ
10              ドキュメント
10              コントロール
10              エキスパート
        （後略）
```

図１：分野総単語リスト

2.3　分野特定単語リストの作成

分野特定単語リストは、分野総単語リストに比べて分野の特徴をより表した単語リストである。分野特定単語リストの作成方法は、分野総単語リストのそれぞれの単語について、分野 i の単語リスト中の単語 j の出現率を求め、このどれかが大きいものをその分野の単語として選択する。分野 i の文献における単語 j の出現率 f_j^i は以下の式によって求める。

$$f_j^i = \frac{N_j^i}{\sum_{j \text{ in } i} N_j^i}$$

ここで、N_j^i は分野 i の単語 j の出現頻度である。また、分野 i にその単語がない場合は、出現率の比が求められないので小さい値 0.1 を与えた。

次に、その比 R は以下の式によって計算する。

$$R_{i_1 i_2 j} = \frac{f_j^{i_1}}{f_j^{i_2}}$$

分野特定単語リストの選択基準は、$R_{i_1 i_2 j} > m$ なら、i_1 の単語リストに加える。$R_{i_1 i_2 j} < \frac{1}{m}$ なら、

表 1: 分野総単語リスト中の語の出現率平均 (%)

単語リスト	文献の出所		
	情報処理学会	日本農芸化学会	土木学会
情報処理学	94.8	85.7	89.6
農芸化学	88.0	94.8	89.0
土木学	91.3	87.9	94.8

表 2: 分野特定単語リスト中の語の出現率平均 (%)

単語リスト	文献の出所		
	情報処理学会	日本農芸化学会	土木学会
情報処理学	12.9	1.2	4.4
農芸化学	0.9	15.0	3.0
土木学	5.7	1.0	9.8

i_2の単語リストに加える。本研究では、$m = 20$ とした。この場合、1つの分野、または2つの分野だけに1回出現するものはその分野の単語リストに採用されないが、3回以上出現し、他のどれかの分野で0回であれば採用されることになる。

以下に情報処理分野の分野特定単語リストの一部を示す。

（前略）
ツール
ツリー
テーブル
テーマ
テキスト
テストケース
テストプログラム
データベース管理システム
データベース機能
データモデル
データ型
データ検索
データ構造
データ処理
データ通信
データ伝送
データ部
（中略）
ネットワーク
ネットワーク管理
ネットワーク型
ネットワーク表現
ハードウェア
ハザード
ハッシュ
ハッシュ関数
バックトラック
バッチ処理
（中略）
プログラミング
プログラム
プログラム言語
プログラム作成
プログラム変換
プロセス管理
プロセス間通信
プロセス制御
プロセッサ
プロセッサー
（後略）

図 2：分野特定単語リスト

表 3: 分野特定単語リストによる分野判定結果 (%)

文献の出所	分野判定結果		
	情報処理学	農芸化学	土木学
情報処理学会	98.3	0.0	1.7
日本農芸化学会	0.3	99.7	0.0
土木学会	5.7	5.1	89.2

3　検索質問作成手法の分析

3.1　単語リスト中の語の文献における出現率

2節で作成した情報処理、土木、農芸化学の分野の2種類の単語リストが、それぞれ3つの分野の文献ごとにどの程度出現しているかを調べるために、文献中における各リスト中の単語を検索し、その出現率を求めた。検索対象文献は、単語リスト作成材料とした文献セットと同じ学会の同一時期の抄録であるが、すべて異なる文献からなる。出現率 $S_k(i)$ は、次式によって求める。

$$S_{k(i)} = \frac{\sum_{j \text{ in } i} H_k^j}{W_k} \times 100$$

ここで、H_k^j は文献 k 中の単語 j の出現頻度、W_k は文献 k 中の総単語数である。

3.2　単語リストの分析

分野総単語リストと分野特定単語リスト中の語の各分野の文献中での出現率を調べるために、出現率の頻度分布を求めた。

この頻度分布のグラフの一例として、情報処理学会の文献のグラフを図3、図4に示す。

この結果から、分野総単語リストと分野特定単語リストとも、3つの分野の文献が異なる山形の分布となり、単語リストと同じ分野の文献はこの中で比較的高い出現率の山に対応する。

分野総単語リストより、分野特定単語リストの方が山の分離が比較的よいが、これだけでは各分野の分布に重なりがある。

次に、各分野の文献における各分野の単語リスト中の単語の出現率の平均を求めた。i 番目の単語リスト中の語の i' 分野の文献集合における平均出現率 $A_{(i,i')}$ は、次式によって求めた。

$$A_{(i,i')} = \frac{\sum_{k \text{ in } i'} S_{k(i)}}{M_{i'}}$$

ここで、$S_k(i)$ は出現率、$M_{i'}$ は i' 分野の対象文献数である。

その結果を表1、表2に示す。

表1からわかるように分野総単語リストは、「に、は、の」という助詞なども含まれた全単語であるため、分野ごとの差が少ない。それに対し表2からわかるように、分野特定単語リストでは、分野の特徴を表す単語を選択しているため、値は小さいが、分野ごとの差は表1に比べて大きい。

両者ともに、単語リストと同じ分野の文献を対象とした場合には平均値が高くなっていることは、各単語リストが分野ごとの特徴をある程度あらわしていることを意味するが、当然ながら分野特定単語リストの方により特徴が表れている。

最後に、個々の文献について各単語リスト中の語の出現率にどのような違いがあるかを調べるため、出現率の比の頻度分布を求めた。文献 k における単語リスト i 中の語の出現率 $R_k(i)$ を次式で定義する。

$$R_k(i) = \frac{S_k(i)}{S_k(i_0)} \times 100$$

ここに、i_0 は k が属する分野を示す。。

この頻度分布のグラフの一例として、情報処理学の単語リストに関するものを図5、図6に示す。

図5からわかるように、分野総単語リストでは出現率の差が小さく、大部分が 91～100 パーセントの間となっている。これに対し、図6からわかるように、分野特定単語リストでは出現率の比の差が、分野総単語リストに比べて大きい。この分析を見る限り分野特定単語リストの方がより分野の特徴が表れているといえる。

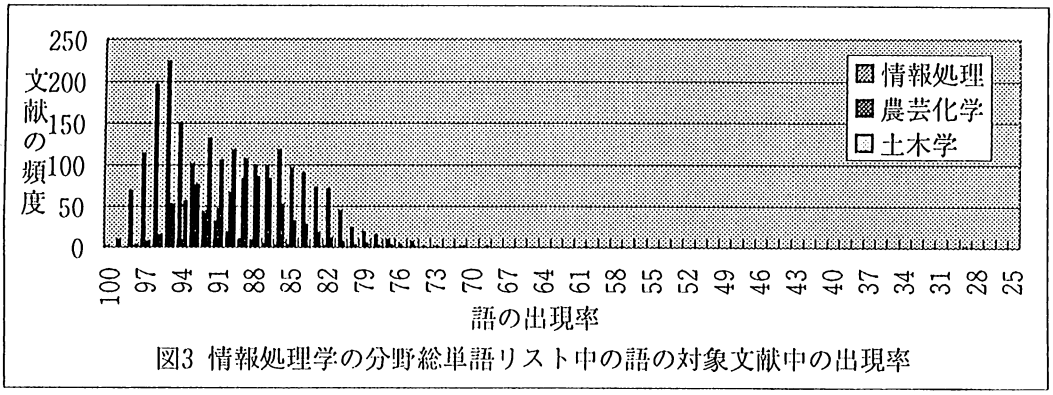

図3 情報処理学の分野総単語リスト中の語の対象文献中の出現率

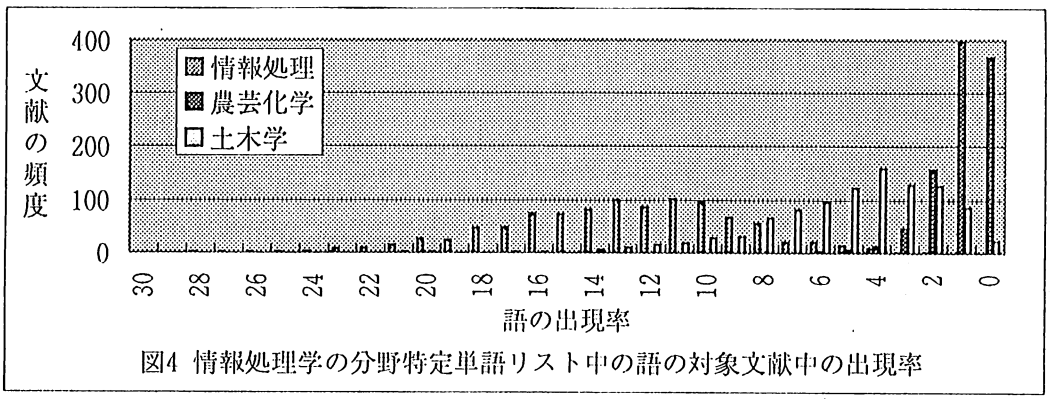

図4 情報処理学の分野特定単語リスト中の語の対象文献中の出現率

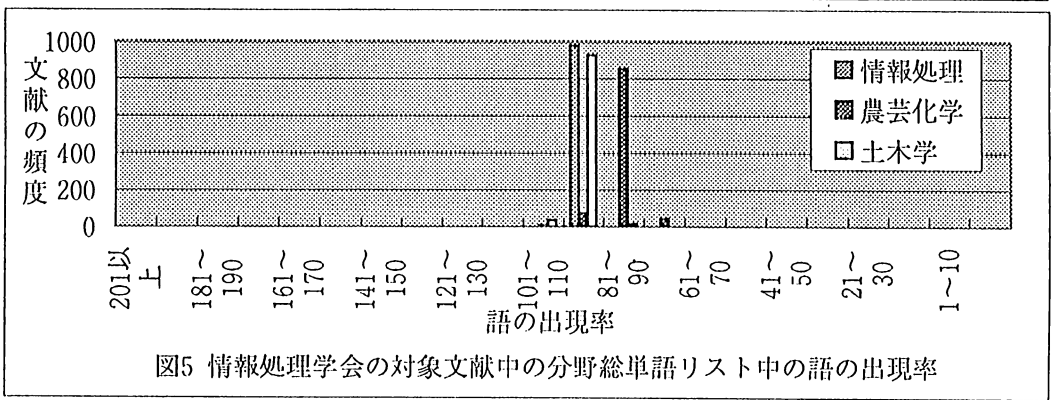

図5 情報処理学会の対象文献中の分野総単語リスト中の語の出現率

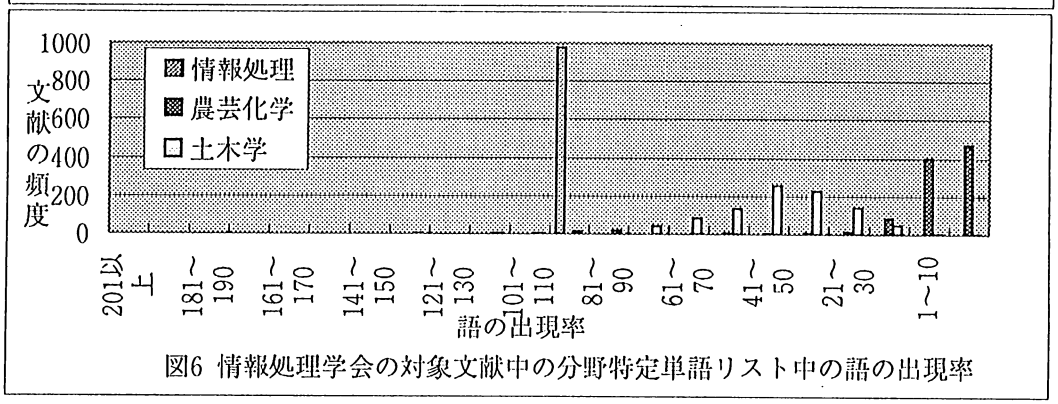

図6 情報処理学会の対象文献中の分野特定単語リスト中の語の出現率

3.3　分野検索法

前節の結果から、分野特定単語リストを用いる分野検索法を次のように定めた。

1. 対象文献に対して、各分野の分野特定単語リスト中の単語の出現率 $S_k(i)$ を求める。
2. $S_k(i)$ の最大値を与える i を、この文献の分野とする。

このようにした場合、各分野の文献がもとの分野に属すると判定された場合、あるいは別の分野と判定された場合がどれほどあるかを表3に示す。表3の対角要素がもとの分野と判定される割合である。対角要素の平均は約96%である。

この表から、上の方法で比較的高い確率で原分野に判定されることがわかる。但し、土木学会の文献のうち、5.7%が情報処理学、5.1%が農芸化学と判定されている。これらについて原情報にあたると、たとえば土木学会の分野の文献で情報処理分野と判定されたものは、

1. 数値解析、数値シミュレーション、データ解析、統計処理などを扱った研究 (30 件)
2. マルチメディア、データベースなどシステムを扱った研究 (12 件)
3. ニューラルネットなどを扱った研究 (6 件)
4. コンテナ管理、流通、輸送などを扱った研究 (8 件)

などであり、4を除いては分野判定の誤りとはいいきれない。

また、同じく農芸化学分野と判定されたものは、

1. 土壌、水質、河川、土壌内微生物などを扱った研究 (17 件)
2. 化合物、化学実験などを扱った研究 (14 件)
3. 材質、コンクリートなどを扱った研究 (14 件)
4. 数値解析、実験装置などを扱った研究 (5 件)

などであり、4を除けば誤りとはいいがたい。

以上のように、上に示した方法で、これら3つの分野の文献の分野判定が可能なことが示される。

4　まとめ

本稿では、情報処理学、農芸化学、土木学の抄録から、複数ハッシュふるい分けによる日本語分かち書きシステム MIISA により切り出した「分野総単語リスト」と、分野の特徴を表すため、語の各分野における出現率の比が、ある程度以上大きい単語を抽出した「分野特定単語リスト」を作成した。

この結果、「分野特定単語リスト」を用いる検索により、上記3分野の抄録をあわせた対象データを 96% 程度の精度で原分野に分離できることがわかった。

参考文献

[1] Salton, G. and McGill, M.J. Introduction to Modern Information Retrieval McGraw-Hill, NewYork, 1983.

[2] 中本賢一. 複数ハッシュふるい分け法の日本語情報システムへの応用. 情報システム研究会. 48-7. 1994.3.

[3] Kigen Hasebe, Ken'ichi Nakamoto, Takeo Yamamoto. An Information Retrieval System on Internet for Languages without Obious Word Delimiters. Proceedings of International Symposium on Digital Libraries 1995. pp.181-185. 1995.8.

石田 栄美　宇陀 則彦　石塚 英弘　山本 毅雄：図書館情報大学
(University of Library and Infomation Science)
根岸 正光：学術情報センター
(National Center for Science Information Systems)