概念体系に基づく情報整理支援ツールの日本語化*

3Y－7

猪股 健太郎　　高間 康史　　石塚 満

東京大学工学部電子情報工学科

1　はじめに

我々の研究室では、インターネットなどを通じて大量に収集された情報を熟読し有効に活用するため、ユーザの視点・興味に従った情報整理を支援するツールを開発している。本稿では、EDR 日本語単語辞書と形態素解析ツール茶筌を組合せ、我々のツールを、日本語の文書を扱えるように拡張したことを報告する。

2　情報整理支援ツールの概要

インターネットに代表される情報環境の急激な整備・拡大により、研究や仕事などに必要とされる情報の収集は容易になってきたが、一方、入手可能な情報量が人間の情報処理能力を超え、かえって効率が低下するという「情報過多」が問題となってきている。例えば、あるテーマや問題に関して調査する為に、サーチエンジンを利用し、関連があると思われる情報 (文書) を収集してみると、短時間で思った以上に多量の文書を集める事ができる。それらの文書を全て読み、内容を理解する事は、労力を要する作業である。

一つ一つの文書を熟読するためには、「今読んだ文書は、さっき読んだあの文書と似ている／対立する」などと、文書間の関係を意識し整理しながら読み進めていく事が重要であると考えられる。我々のツールは、このような情報整理過程を視覚的に支援することで、ユーザの労力を少なくし、各文書をより深く熟読する事ができるようにするものである。

我々のツールは、ユーザの視点に関連した文書の特徴情報を、特徴ベクトルとして表現する。ただし、各単語を直交した特徴として扱うのではなく、概念体系を用いてグループ化することができるようにする。これにより、ユーザの興味・視点に合わせて特徴ベクトルを動的に生成できる。生成された特徴ベクトルを、**Fish Eye** ベクトルと呼ぶ。

インターネットはアメリカで発展したという経緯もあって、インターネット上の多くの情報は英語で書かれている。ところが、日本でもコンピュータの急激な普及によって、一般家庭からプロバイダと契約してインターネットに接続する人が増えた。それによって、インターネット上の日本語の情報も多量になって来た。日本語のみの情報もあるし、元は英語で書かれていた情報で、日本語に訳されたものもある。日本語の情報とその受け手が増えている今、我々のツールを、日本語を扱えるように拡張することは意義があると言える。

3　Fish Eye ベクトル

基本特徴ベクトルに以下の 2 種類の操作を施すことによってユーザの視点・興味に沿った特徴ベクトルに作り直す。

- **縮退 (shrink)**
  指定した概念グループに属する単語を全て同じ特徴として扱う
- **拡大 (magnify)**
  指定した概念グループ集合のいずれかに属する単語のみを特徴として扱い、他は切り捨てる

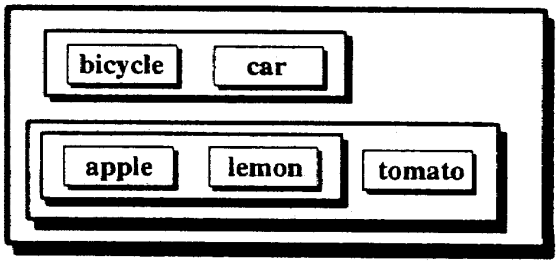

図 1：概念グループの例

---

* Improvement of handling Japanese of Information arrangement support tool based on Concept Structure.
Kentaro Inomata, Yasufumi Takama, Mitsuru Ishizuka
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
inomata@miv.t.u-tokyo.ac.jp

図1の例では、“食べ物”という概念に着目すれば(magnify)、bicycle や car といった単語は無視される。“食べ物”のうち、apple と lemon は同じ“果物”なので区別しないことにする(shrink)こともできる。

4　ツールの日本語化

我々のツールは、以下の4つの処理から成る。

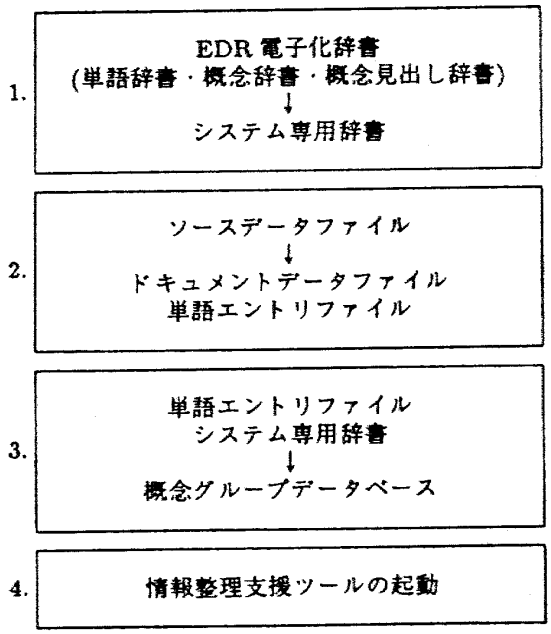

図2：ツールの構成

我々のツールでは、概念体系として、EDR 概念体系辞書を用いている。この辞書を選ぶ利点は、概念体系が英語・日本語で共通であるので、ツールの日本語化において概念辞書を操作する部分はほとんど変更せずに済むことである。

ソースデータファイルからドキュメントデータファイル・単語エントリファイルを作るプログラムでは、文書中の単語を切り出す必要がある。英語の文書の場合は単語と単語はスペースで区切られているので計算機での処理を行うことはたやすいが、日本語の文書は全ての単語が一続きであり、単語を切り出すためには形態素解析を行う必要がある。その処理には、形態素解析ツール「茶筌」のライブラリを用いた。

5　茶筌と EDR 単語辞書

茶筌はシステム標準の辞書を備えている。しかし、茶筌で切り出す単語を EDR 概念体系辞書と関係付けるためには、茶筌用辞書の単語エントリと EDR 日本語単語辞書の単語エントリが一致しなければならない。そのため、EDR 日本語単語辞書の情報を茶筌用に再構成した。EDR 辞書と茶筌用辞書は以下のような構成である。

EDR 日本語単語辞書の構成：

```
レコード番号 \t 単語見出し \t 不変化部-連接
属性対 \t かな表記 \t 発音 \t 品詞 \t 構文木
\t 活用情報 \t 表層格情報 \t 相情報 \t 機能語
情報 \t 概念識別子 \t 英語概念見出し \t 日本
語概念見出し \t 英語概念説明 \t 日本語概念説
明 \t 用法 \t 頻度 \t 管理情報 \n
```

茶筌用形態素辞書の記述例：

```
(名詞
  (普通名詞
    ((見出し語 日本語)
     (読み にほんご))
  )
)
```

現在我々のツールでは、簡単のため名詞のみを特徴としているので、EDR 日本語単語辞書・日本語専門用語辞書から名詞のみを取り出し、茶筌用に再構成した。また、名詞の細分類も EDR のものを利用した。その他の品詞については茶筅に付属の辞書を使用している。

EDR 辞書の単語見出しは“日本語 [ニホンゴ]”のように漢字混じりの標記とカナ標記が一体となっているので、これを切り分けて茶筌の見出し語情報と読み情報に対応させた。

6　おわりに

日本語化に際し、辞書を日本語辞書のみにしたので、英単語や、アルファベットの固有名詞を特徴として扱えない。今後は英語・日本語混じりの文を扱えるように改良したい。また、現在、特徴として扱う単語は名詞のみである。用言の活用に対する扱い方がが EDR 日本語単語辞書と茶筌で大きく異なるので問題もあるが、今後は動詞や形容詞・副詞なども特徴として扱えるよう改良したい。

参考文献

[1] 高間, 石塚：概念体系を用いた Fish Eye ベクトルの情報整理支援ツールへの応用, 人知研資 SIG-FAI-9702, pp. 97–102(1997).

[2] http://www.iijnet.or.jp/edr/J_index.html

[3] http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html