# 想起型情報検索システムにおける文書のクラスタ化

1V－8

熊本 睦 *　飯田 敏幸 **

*NTT コミュニケーション科学研究所　**NTT 第一法人営業本部

## 1 はじめに

従来から多く利用されているキーワード(KW)一致による情報検索システムは、探したい対象がほぼはっきりして、それに関するKWが分かっていることを前提としている。このため、探したい対象が漠然としている場合は余り役に立たない。

我々は、このような状況にでも対応できることを目標として、想起型情報検索システムを試作した[1]。本システムでは、検索対象である新聞記事をクラスタ分けし、各クラスタの特徴語を提示することにより、利用者が必要なKWを見い出し、検索要求を明確化することを支援する。

## 2 検索要求の明確化支援

検索意図が明確でない場合、ただ漫然と情報を見ていても有用な情報は得られない。検索システムはヒントになる情報を積極的に検索者に与える必要がある。本システムでは、内容が似ている情報同士をクラスタとしてまとめ、各クラスタがどのような情報の集まりであるかを表す特徴語を提示することにより、検索要求の明確化を支援する。利用者は、最初にKWが思いつかない場合でも、提示された特徴語を見て、関心のあるクラスタを選択したり、特徴語からKWを選択したり、特徴語を見て思いついたKWを指定して検索することができる。

## 3 文書のクラスタ化とクラスタ特徴語

検索要求の明確化支援のうち、クラスタ化とその特徴語の決定方法について述べる。

### 3.1 クラスタ特徴語

本システムでは、概念ベース(単語をベクトルで表現した知識)に基づき、単語同士の類似度を単語に対応するベクトルの内積で定義している。概念ベースは2種類作成している。1つは、辞書に基づく概念ベース(辞書CB)で、国語辞書の説明文中の単語頻度を元にベクトルを作成している[2]。もう1つは、新聞記事に基づく概念ベース(記事CB)で、記事中の単語の共起頻度を元にベクトルを作成している[3]。

文章を単語の羅列と捉え、文章中の単語に対応するベクトルの平均を文章のベクトルとして定義する。また、文章と文章の類似度は対応するベクトルの内積で定義する。クラスタ分けにはBuckshotアルゴリズム[4]を用いている。また、クラスタの特徴語は、そのクラスタに属する記事に対応するベクトルを平均したベクトル(クラスタ重心ベクトル)との類似度が高いベクトルに対応する単語とする。

### 3.2 クラスタ分けの例

**例1)** 利用者が探したい事柄を全く決めていない場合を考える。この場合は、ある指定された期間の記事を対象としてクラスタ分けし、特徴語を利用者に提示する。例として、日本経済新聞[1]1995年2月1日の記事718件の見出しを対象にクラスタ分けし、特徴語を求めた結果を表1、2に示す。

**例2)** 利用者が漠然とある事柄について探したいが、その事柄についてあまり分かっていない場合を考える。この場合は、最初に適当なKWを使って検索し、その結果をクラスタ分けし、利用者に特徴語を提示する。例として、通信と関連した何かの情報を探したいという場合を示す。日本経済新聞1995年2月の記事を対象に「通信」で検索し、類似度上位500記事についてその見出しを対象にクラスタ分けし、特徴語を求めた結果を表3、4に示す。

### 3.3 考察

クラスタ分けの結果(表1～4)はCBにより異なっているが、クラスタ中の記事をチェックしたところ、どちらのCBによる場合も意味のあるクラスタになっていることが分かった。また、提示される特徴語には、次のような傾向がある。

---

Document Clustering in Associative Information Retrieval System

*Mutsumi Kumamoto
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
**Toshiyuki Iida
NTT Business Communications Headquarters I
1-1-6 Uchisaiwai-cho, Chiyoda-Ku, Tokyo, 100-8019 Japan

[1] 日本経済新聞 CD-ROM1995 年版を使用

辞書CBによるクラスタ分けの場合、特徴語として類義語が並ぶ。例えば、表1の阪神大震災に関する記事は、「災害」、「被災」、「震災」のようになっている。これは、辞書CBでは国語辞書の説明文からベクトルを作成しているため、類義語になる単語同士の類似度が高くなるからだと考えられる。

一方、記事CBによるクラスタ分けの場合、各特徴語の意味の共通項がクラスタ内の記事の傾向を示している。例えば、表4で、「電話」、「NTT」、「携帯」、「PHS」の意味の共通項を考えると、それは移動体通信を示していると言える。これは、記事CBの場合、共起する単語によりベクトルを定義しているため、共通項になるような単語と共起する単語の類似度が高くなるからだと考えられる。

関連研究として、Cutting et al.[4]の研究がある。彼らの方法では、文書を文書中の単語頻度のベクトルで表し、クラスタ中の高頻度の単語を特徴語としているので、特徴語はクラスタ中の記事の単語からしか選ばれない。これに対し、我々の方法ではクラスタ中の記事にない単語でも、重心ベクトルとの類似度が高ければ特徴語となる。従って、新しい発想を得たいような場合に向いていると考えている。

## 4 おわりに

想起型情報検索システムにおける文書のクラスタ化および利用者に提示するクラスタ特徴語について述べ、クラスタ分けの例を示した。今回、使用したBuckshotアルゴリズムは、あらかじめクラスタ数を指定する必要があり、クラスタ数によっては適切なクラスタ分けにならない。そこで、適切なクラスタ数を自動的に決める方法を検討する予定である。

想起型情報検索システムの研究開発はIPA創造的ソフトウェア育成事業による。

## 参考文献

[1] 飯田他: 想起型情報検索システムについて, 情処研究報告, 98-OS-77/98-DPS-87, pp.19–24 (1998).

[2] 笠原他: 国語辞書を利用した日常語の類似性判別, 情処論文誌, Vol. 38, No. 7, pp.1272-1284 (1997).

[3] Schütze et al.:Information Retrieval Based on Word Senses, 4th Annual Symposium on Document Analysis and Information Retrieval, pp.161–176(1995).

[4] Cutting et al.: Scatter/Gather: A Cluster-based Approach to Browsing Large Documant Collections, SIGIR '92, pp.318–329(1992).

表 1: 辞書CBによるクラスタ分け (1日分)

| クラスタ特徴語 * | 記事の傾向 |
|---|---|
| 災害、被災、震災、罹災 | 阪神大震災 |
| 相場、単価、値、価、急落 | 株式、相場 |
| 親会社、会社、社業、事業 | 企業の状況 |
| 令書、訴訟、請求、抗告 | 裁判 |
| 死去、物故、永眠、逝去 | 死亡記事 |
| 開始、店開き、開幕 | 営業、生産の開始 |
| 相談、商議、話し合い、協議 | 会合 |

表 2: 記事CBによるクラスタ分け (1日分)

| クラスタ特徴語 * | 記事の傾向 |
|---|---|
| 大震災、阪神、被災、兵庫県 | 阪神大震災 |
| 開通、運行、特急、区間、雪 | 交通 |
| 起訴、同意、裁、違反、異議 | 刑事、民事事件 |
| 伊藤忠、合弁、持ち株、解禁 | 企業の活動 |
| 今期、連結、好調、減益 | 企業の収益 |
| 急落、反落、続伸、小幅、相場 | 金融取引 |
| 金融市場、日銀、ＴＢ、オペ | 金利 |
| 北陸、栃木、茨城、静岡、新潟 | 地方 |

表 3: 辞書CBによるクラスタ分け (通信)

| クラスタ特徴語 * | 記事の傾向 |
|---|---|
| 通信、着信、発信、入電 | 情報通信 |
| 事業、親会社、企業、資本 | 通信事業 |
| 災害、被災、罹災、震災 | 通信システムの被害 |
| 機器、器機、器械、機械 | 通信機器 |
| 電話、通話、電話機、混線 | 電話 |
| 緩和、規制、雪解、規則 | 通信の規制緩和 |

表 4: 記事CBによるクラスタ分け (通信)

| クラスタ特徴語 * | 記事の傾向 |
|---|---|
| 情報、ネットワーク、発信 | 情報通信 |
| 電話、ＮＴＴ、携帯、ＰＨＳ | 移動体通信 |
| テレコム、通信サービス | 通信事業 |
| テレビ、放送、有線、ラジオ | 放送 |
| 規制、緩和、撤廃、小委、参入 | 通信の規制緩和 |
| パソコン、ＮＥＣ、周辺機器 | パソコン通信 |

* クラスタ重心ベクトルとの類似度順