

## 特許検索における分類手法の比較評価

牧田 光晴<sup>†</sup> 樋口 重人<sup>†</sup> 藤井 敦<sup>†,††</sup> 石川 徹也<sup>††</sup>

<sup>†</sup> (株)パトリス

〒135-0043 東京都江東区塩浜 2-4-29

<sup>††</sup> 図書館情報大学

〒305-8550 つくば市春日 1-2

<sup>†††</sup> 科学技術振興事業団 CREST

E-mail: m\_makita@patolis.co.jp

あらまし 特許検索において、膨大な検索結果を組織化することが近年重要な課題となっている。1つの解決法としてクラスタリングがある。しかしクラスタリングの定量的な評価はこれまでほとんど行われていない。そこで筆者らは、適合文書がクラスタに含まれている確率をもとにエントロピーの概念を導入した評価手法を提案する。また、テストコレクション中の適合の度合の情報を採り入れた評価を行う。さらにクラスタのサイズに関する実験を行い、クラスタリングについての多方向からの評価を試みることにより、クラスタリング手法の妥当性を調査する。

## Comparative Evaluation of Document Clustering Methods in Patent Retrieval

Mitsuharu Makita<sup>†</sup>, Shigeto Higuchi<sup>†</sup>, Atsushi Fujii<sup>†,††</sup>, Tetsuya Ishikawa<sup>††</sup>

<sup>†</sup>PATOLIS Corporation

2-4-29 Shiohama Koto-ku, 135-0043, Japan

<sup>††</sup>University of Library and Information Science

1-2 Kasuga Tsukuba, 305-8550, Japan

<sup>†††</sup>CREST, Japan Science and Technology Corporation

E-mail: m\_makita@patolis.co.jp

**Abstract** Recently, in the domain of patent information retrieval, it becomes of great importance to organize enormous amounts of retrieved documents. Although document clustering is one of effective methods in organizing documents, few attempts have been made to evaluate clustering methods quantitatively. In this paper, we propose an evaluation method for document clustering, which is based on the concept of entropy (i.e., uncertainty). Our evaluation method can be used for experiments where relevance assessment is performed based on multi-level relevance. We used our evaluation method to perform several comparative experiments and investigated the effectiveness of existing document clustering methods from a number of different perspectives.

## 1 はじめに

近年、特許検索の領域において海外特許検索のニーズが高まっている。これに応じ、筆者らは多言語特許検索システム PRIME (Patent Retrieval In Multi-lingual Environment) を開発した [6]。PRIME は、検索キーワードを検索対象の言語に翻訳して外国語特許を検索し、検索された特許を母国語に翻訳することで多言語検索を実現する。翻訳と検索の評価については既に行った評価実験によりおおむね良好な結果を得ている。

また、膨大な検索結果を組織立ててユーザに提示することが盛んに求められていることを受け、PRIME では、検索結果にクラスタリングを施し、閲覧効率を向上させることを可能にした [7]。これまでに、クラスタリングについての定量的な評価はほとんど行われてこなかったため、ユーザの主観に依存した形での評価を超えるものは少なかった。本稿では、クラスタリングの客観的な評価を可能とする評価尺度を提案し、種々のクラスタリング手法の評価を行う。

以下、2章でクラスタリングと提案する評価手法について説明し、3章で PRIME の概要を説明し、4章と5章では実験の説明を行い、6章で考察を行い、7章でまとめと今後の研究課題について述べる。

## 2 検索結果の分類とその評価手法

### 2.1 カテゴリゼーションとクラスタリングの相違点

文書を効率的に組織化する手段には幾つか手法がある。1つは、トップダウンに文書に分類を付与方法 (カテゴリゼーション) である。医学や特許などの専門分野では高度に発達した分類を利用した文書検索を行う場合が多く、これによって、検索だけではなく検索された文書の組織化もスムーズに行えるようになっている。例えば、特許の領域では、国際特許分類 (IPC)、米国特許分類、FI (File Index: 日本独自の特許分類) 等の幾つかの分類が用途ごとに使い分けられている。

しかし、これらの分類体系は非常に複雑であるため、これらを駆使することは、ユーザにとって多大な負担となる。このため、検索結果の表示の際に、ユーザが分類を用いることなくボトムアップに文書の集合を作成することが求められる。これが文書の組織化の2つ目の手段、すなわちクラスタリングである。

本研究で、クラスタリングのエンジンに GETA [11] を使用した。GETA は、大規模文書の高速な連想計算が可能なソフトウェアである。GETA はクラスタリング用のライブラリを持ち、5つのクラスタリングアルゴリズムが使用可能である。本稿では、この5つのアルゴリズムを比較する。

### 2.2 クラスタリングの評価

クラスタリングの評価法は、僅かながら存在する。Hearst ら [5] は、まず検索結果をクラスタリングして作成された複数のクラスタから、クラスタ中の適合文書数を元に唯一の「最適な」クラスタを自動的に決定し、適合文書 (正解) が最適クラスタに含まれる確率によって、クラスタリングの精度を評価した。

Fujii らは [2]、クラスタリングの評価にエントロピーの概念を導入した。すなわち、特定のクラスタに適合文書が集中し、適合文書の分布に偏りが生じるほど、良いクラスタリング手法と見なし、逆にどのクラスタにも適合文書が一樣に分布する場合には、ユーザが適合文書を選択するための情報がほとんど無いので、クラスタリングの精度は低いと見なす手法である。

筆者らは、この手法を用いて多言語特許検索におけるクラスタリングの評価を試みた [10]。

### 2.3 提案手法

ここではまず文書の「適合レベル」について論じる。これまでのテストコレクションでは、適合/不適合の2値の尺度で正解文書が決定されることが多かった。しかし最近 { 完全に適合, 部分的に適合, 不適合 } のような3つ以上の尺度による適合の度合 (以下適合レベルと呼ぶ) に基づいて正解文書を定めているテストコレクションが存在する。本稿ではこの3つ以上の適合レベルを扱うことができる一般的な評価尺度を提案する。

前述の Fujii らの手法は、1回のクラスタリングでできた2値の適合レベルによる適合文書の散らばりにもとづいたエントロピーを評価尺度とするものであった (式1, 2)。

$$H(X) = - \sum_{c=1}^N P(c) \log P(c) \quad (1)$$

$$P(c) = \frac{\text{クラスタ } c \text{ の適合文書数}}{\text{クラスタ } c \text{ の総文書数}} \quad (2)$$

$H(X)$  はクラスタ全体のエントロピーを、 $P(c)$  はクラスタ  $c$  に適合文書が存在する確率、 $N$  はクラスタの数を表す。

ここで、前述の適合レベルを考慮した評価を行うため、(1),(2) を改良する。まず、 $P(c)$  では適合・不適合の2値に関する確率しか扱えないため、これを以下の  $P_c(r)$  に変更する。

$$P_c(r) = \frac{\text{クラスタ } c \text{ の適合レベル } r \text{ の文書数}}{\text{クラスタ } c \text{ の総文書数}} \quad (3)$$

次に、1つのクラスタについてエントロピーを仮定する。つまり、1クラスタ内の適合レベルごとの文書の散

らばりを基に評価を行う。このとき、適合レベルごとによろしくクラスタリングが行われているほど、1 クラスタ内に異なる適合レベルの文書どうしが共存することは少なくなる (1 クラスタのエントロピーは小さくなる)。

式 (4) に 1 クラスタのエントロピー  $H(c)$  を定義する。

$$H(c) = - \sum_{r=0}^n P_c(r) \log P_c(r) \quad (4)$$

ここで、 $r$ 、 $n$  はそれぞれ、適合レベル、適合レベルの数、である。例えば、適合レベルが適合と不適合のみならば、 $n = 1$  である。

さらに、クラスタリングされた文書全体に対して、適合文書の散らばりに関する情報量を、各クラスタのエントロピーを元に計算する。すなわち、クラスタサイズの重みパラメータ  $\alpha_c$  を用いて各クラスタのエントロピーの加重平均を求めこれを情報量とする。

式 (5) に 1 回のクラスタリングによって作成される文書の散らばりに関する情報量  $I$  を定義する。

$$I = \sum_{c=1}^N \alpha_c H(c) \quad (5)$$

ここで、 $N$  は作られたクラスタの数、 $P_c(r)$  はクラスタ  $c$  に適合レベル  $r$  の文書が含まれる確率を、 $\alpha_c$  はクラスタ  $c$  のサイズの重みをそれぞれ表す。本稿ではこの情報量  $I$  をクラスタリングの評価尺度とすることを提案する。

### 3 PRIME の概要

PRIME のシステム構成を図 1 に示す。実線はオンライン処理、破線はオフライン処理をそれぞれ表す。現在、日本語と英語による検索が可能である。

まず、検索質問翻訳部では、ユーザがキーワードやフレーズなどで検索質問を入力すると、翻訳部において目的言語に翻訳される。ここでは、対訳辞書を引き、単語や複合語の単位での翻訳を実現する。しかし、一つの語に対し複数の訳が存在し、訳語候補の全てを採用するとノイズが増大するため、対訳辞書から抽出した翻訳モデルと検索対象の特許 DB から抽出した言語モデル (単語バイグラム) を使用して統計的手法で翻訳の曖昧性を解消する。我々の先行研究 [3, 7] では、翻訳の精度は、異表記や省略等による表記の違いも許容すれば、日英で 94.2 %、英日で 92.8 % であり、実用レベルに達している。

次に、検索質問と特許の類似度を確率的なスコアで計算し、スコアに基づいて検索された特許をソートしてユーザに提示する。この段階で、検索結果をクラスタリングして、クラスタ (グループ) の単位でユーザに特許を提示することが可能である。その結果、個々の特許ではなく、クラスタ単位で、必要な情報を選択することができる。さらに、この処理を対話的に繰り返しながら、ユーザを適切な情報に効率良く誘導することができる。

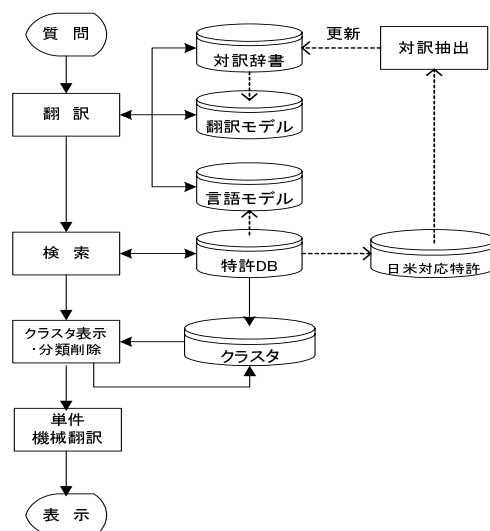


図 1: PRIME のシステム構成

その後、ユーザがクラスタから特許を単件指定すると、その内容がユーザ言語に機械翻訳されて表示される。

また、オフライン処理ではパラレルコーパスから日英対訳を自動抽出し、対訳辞書を随時更新する。

検索対象データは日本特許抄録 1995-1999 年の 5 年分 (約 175 万件)<sup>1</sup>、英語抄録 PAJ (Patent Abstracts of Japan)<sup>2</sup> の同期間、同件数含むパラレルコーパスである。これに加え、優先権出願<sup>3</sup>された日本公開公報全文と、その元となる米国特許登録公報全文の組を約 32,000 件を搭載する。

さらに、PRIME では Scatter/Gather [1, 4, 5] のような、複数回のクラスタリング機能を追加した。これは、 $n$  回目のクラスタリングで作成されたクラスタのうち、ユーザが必要と判断したクラスタのみを選択し、 $n + 1$  回目のクラスタリングを実行するものである。また、1 回のクラスタリング毎に、作成するクラスタの数を変更することも可能である。

<sup>1</sup>(株) パトリスより販売されている。

<sup>2</sup>特許庁より販売されている。

<sup>3</sup>国際条約に基づいた出願形態の一つ。本稿で用いたデータは、米国で出願された特許の権利が日本でも優先的に主張されたものである。

## 4 準備

### 4.1 テストコレクション

本稿では(株)パトリス作成のテストコレクションを使用した<sup>4</sup>。このテストコレクションは、検索質問と正解文書集合を34組含み、さらに正解文書には「発明情報に完全にマッチ(HA)」、「発明情報に部分的にマッチ(HB)」、「非発明情報に完全にマッチ(IA)」、「非発明情報に部分的にマッチ(IB)」の4つの適合レベルが設けられている。

以下にテストコレクションの一部を例示する。

(6-Q) ガス交換にすぐれた中空系膜において、特に血液との間でガス交換性に優れ、酸素ガスと炭酸ガスの交換をおこなう人工肺に好適な中空系膜の材料及び製造方法に関するもの。

(6-HOSOKU) 人工肺用中空系膜の、製造方法、加工方法、処理方法、材料についての発明を対象とする。人工肺モジュールそのものは対象としない。

(6-SEIKAI)28

(6-A) 1998024103HA

(6-A) 1998028727HA

(6-A) 1998036474HB

(6-A) 1998099426HB

(6-A) 1998099652HB

(6-A) 1998109023HB

(6-A) 1998118176HB

...

この例で、(6-Q)は6番目の質問を、(6-HOSOKU)はその補足説明(発明情報と非発明情報に関する等)を、(6-SEIKAI)は正解文書数を、(6-A)は正解文書の文書IDと適合レベルをそれぞれ表す。

ここで、適合レベルの序列は、

$$HA > HB > IA > IB$$

である。

また、適合文書(抄録)の1例を以下に示す。

```
<ID>1998024103</ID>
<TITLE>膜型人工肺</TITLE>
<ABSTRACT><P>( J ) 中空系のガス接触面と血液接触面を親水
化処理することにより、中空系内面に水が滞留してガス交換性
能が低下しないようにする。ε水滴、プライミング、ポリオレ
フィン系ポリマー、ハミルトン、変法</P>
<P>中空系の中空部にガスを流し、中空系の外部に血液を流す
外部灌流型の膜型人工肺において、中空系のガス接触面および
血液接触面を親水化処理する。ここで、血液接触面を、さらに
抗血栓性材料で処理し、また、抗血栓性材料をヘパリンとする。
この膜型人工肺は血液中の水分が水蒸気として膜を通過しても、
中空系内部で凝集して滞留することがない。したがって、長時間
使用してもガス交換性能の低下をきたさず、極めて有用性が
高い。</P></ABSTRACT>
```

<sup>4</sup>同テストコレクションは、NTCIR-3 特許タスクで「パトリス標準検索課題」として配布予定である [8]。

このテストコレクションの検索質問文(捕捉説明は含まない)を形態素解析システム Chasen を用いて形態素解析し、検索タームに成形した。

### 4.2 クラスタリングアルゴリズム

本研究では、GETA に含まれている以下の5つのクラスタリングアルゴリズムと1つの対照法を用いる。

- 単一リンク法 (Single LINKage method; 以下 SLINK) : 2つのクラスタの文書間の類似度の最小値をクラスタの類似度とする。
- 完全リンク法 (Complete LINKage method; 以下 CLINK) : 2つのクラスタの文書間の類似度の最大値をクラスタの類似度とする。
- 群平均法 (Groupe AVErage method; 以下 GAVE) : 2つのクラスタの文書間の類似度の算術平均をクラスタの類似度とする。
- WARD 法 (WARD's method; 以下 WARD) : クラスタの中心からの偏差値をもとに、ある文書があるクラスタに属するか否かを決定する。
- 確率的階層ベイジアン法 (Hierarchical Bayesian Clustering; 以下 HBC[9]) : 文書  $d$ 、文書集合  $C$  において  $P(C|d)$  を用いて、クラスタ内の各文書が自身を含むクラスタを発見することができる確率である自己再現率をクラスタリングの尺度とする手法。
- 対照法 : 高度なクラスタリングは行わずに、検索結果を、クラスタリングを行った場合と同じ数の均等なサイズの文書群に、検索ランク順に分割する(以下 BASE)。

## 5 評価実験

本章では検索結果のクラスタリングについて行った3つの実験について述べる。

実験1では、式4、式5の尺度を用いて、適合・不適合の2値の適合レベルによる評価を行う。実験2では、実験1と同じ尺度を用いて、5段階の適合レベルによる評価を行う。また、適合文書が存在するクラスタのサイズが極端に大きくなる場合があり、このことがクラスタリングの評価に与える影響を調べるため、実験3ではCLINKアルゴリズムによるクラスタリングを複数回行ったときのクラスタのサイズに関する実験を行う。

実験1,2では、34検索課題について検索し、検索結果のクラスタリングの処理を、6種類のアルゴリズム、3種類の最大検索文書数についてそれぞれ行う。よって各実験においてクラスタリングを612回ずつ実行することになる。また、クラスタサイズの式(5)中の重みパラメータ  $\alpha_c$  は、以下の実験で  $s(c)$ 、 $\sqrt{s(c)}$ 、1の3つの異なる値を用いる。 $s(c)$  は、式(6)で示される。

$$s(c) = \frac{\text{クラスタ } c \text{ の文書数}}{\sum_{c=1}^N \text{クラスタ } c \text{ の文書数}} \quad (6)$$

## 5.1 実験 1

### 5.1.1 実験条件

以下に実験条件を示す.

- アルゴリズム: {SLINK, CLINK, HBC, WARD, GAVE, BASE},
- 作成するクラスタ数: 5
- 最大検索文書数: {100, 500, 1000}
- 対象言語: 日本語
- $\alpha_c: \{s(c), \sqrt{s(c)}, 1\}$
- 適合レベル:  $(0 \leq n \leq 1)$

ここで最大検索文書数は, 検索時にシステムが返す文書数の最大値で, 本研究のすべての実験において, 検索ヒット数が最大検索文書数を上回ったため<sup>5</sup>, この値は 1 回目クラスタリングを行う文書数になる.

### 5.1.2 結果

最大検索文書数 100, 500, 1000 の場合における, 各アルゴリズムの情報量の平均を, それぞれ表 1~表 3 に示す.

この表で, 情報量の値が低いほど優れた手法と見なす. また,  $\Delta I = (\text{BASE の情報量}) - (\text{各クラスタリングアルゴリズムの情報量})$  が正であるときクラスタリングによって適合文書が得やすくなることを, 負のときは適合文書が得にくくなることをそれぞれ表す.

表 1~表 3 の大半の場合において, 最大検索数が増すにつれて  $\Delta I$  の値は減少している. これはクラスタリングする文書数が少ない程, クラスタリングによって適合文書が得やすくなっていることを示している.

また, クラスタサイズのパラメータ  $\alpha_c$  別で見ると, HBC の場合,  $\alpha_c = s(c)$  では他のアルゴリズムに比べてあまり変わらない. しかし,  $\alpha_c = 1$  ではこれが極端に悪い値となっている.

## 5.2 実験 2

実験 2 では, テストコレクションに記載されている適合レベルの情報を考慮した評価を行う.

<sup>5</sup>これは比較的大きな検索質問語数 (複合語を 1 語と見なせば平均 8.21 語) で or 検索を行っていることが影響していると考えられる.

表 1: 各アルゴリズムの情報量の平均 (実験 1, 最大検索文書数 100)

$\alpha_c$	BASE	CLINK	GAVE	HBC	SLINK	WARD
$s(c)$	0.424	0.442	0.423	0.416	0.393	0.409
$\sqrt{s(c)}$	0.949	0.546	0.681	0.915	0.791	0.700
1	2.12	0.987	1.51	2.01	1.75	1.53

表 2: 各アルゴリズムの情報量の平均 (実験 1, 最大検索文書数 500)

$\alpha_c$	BASE	CLINK	GAVE	HBC	SLINK	WARD
$s(c)$	0.239	0.266	0.256	0.239	0.245	0.239
$\sqrt{s(c)}$	0.533	0.275	0.307	0.512	0.424	0.439
1	1.19	0.375	0.479	1.12	0.864	1.02

### 5.2.1 実験条件

ここでは, テストコレクションの適合レベル  $n$  の値を代える. すなわち実験 1 では,  $n$  は,

$$0 \leq n \leq 1 \\ (n = 0(\text{不適合}), n = 1(\text{適合}))$$

の 2 値であったところを, 実験 2 においては,

$$0 \leq n \leq 4 \\ (n = 0(\text{不適合}), n = 1(HA), n = 2(HB), n = 3(IA), n = 4(IB))$$

を用いる.

以上の点以外は実験 1 と同一の条件の下で実験を行う.

### 5.2.2 結果

最大検索文書数 100, 500, 1000 の場合における, BASE の 34 検索課題の情報量の平均と各アルゴリズムの情報量の平均の差を, 表 3~表 5 に示す.

全体的に数値が若干上がる点を除けば, 実験 1 とほぼ同様の結果であった. これは, 1 クラスタ内の適合レベ

表 3: 各アルゴリズムの情報量の平均 (実験 1, 最大検索文書数 1000)

$\alpha_c$	BASE	CLINK	GAVE	HBC	SLINK	WARD
$s(c)$	0.168	0.190	0.183	0.174	0.179	0.170
$\sqrt{s(c)}$	0.375	0.196	0.213	0.375	0.258	0.322
1	0.839	0.324	0.453	0.835	0.550	0.796

表 4: 各アルゴリズムの情報量の平均 (実験 2, 最大検索文書数 100)

$\alpha_c$	BASE	CLINK	GAVE	HBC	SLINK	WARD
$s(c)$	0.484	0.525	0.495	0.470	0.450	0.473
$\sqrt{s(c)}$	1.08	0.629	0.773	1.03	0.900	0.798
1	2.42	1.04	1.62	2.32	1.96	1.69

表 5: 各アルゴリズムの情報量の平均 (実験 2, 最大検索文書数 500)

$\alpha_c$	BASE	CLINK	GAVE	HBC	SLINK	WARD
$s(c)$	0.272	0.311	0.297	0.270	0.283	0.273
$\sqrt{s(c)}$	0.607	0.317	0.352	0.578	0.480	0.498
1	1.36	0.376	0.501	1.26	0.95	1.14

ルの散らばりによる影響が値の増加として表れていることと、5 値の評価においてもアルゴリズムごとの乱雑さの違いは 2 値の場合と変わらないことを示す。

### 5.3 実験 3

実験 3 では、クラスタリングのサイズに関する実験を行う。

クラスタリングアルゴリズムによって、作成されるクラスタの大きさは大きく異なる。例えば、同じ条件の下、HBC では個々のクラスタのサイズはほぼ均等であるのに対し、CLINK や SLINK では、ある一つのクラスタのサイズが他のクラスタと比べて非常に大きくなることが多い。

以上のことを詳細に調査するため、ここで、一つのクラスタのサイズだけが極端に大きくなりやすい CLINK において、反復的にクラスタリングを繰り返した時のクラスタのサイズと、適合文書の散らばりを計る。

表 6: 各アルゴリズムの情報量の平均 (実験 2, 最大検索文書数 1000)

$\alpha_c$	BASE	CLINK	GAVE	HBC	SLINK	WARD
$s(c)$	0.189	0.218	0.210	0.194	0.205	0.193
$\sqrt{s(c)}$	0.422	0.223	0.240	0.419	0.289	0.364
1	0.944	0.308	0.455	0.930	0.577	0.890

表 7: クラスタサイズと適合文書数 (CLINK)

クラスタリング回数	1	2	3	4	5
平均クラスタサイズ	926.3	858.8	794.2	657.5	411
平均適合文書数	20.0	20.1	12.6	3.00	1.00
質問の数	15	9	5	3	1

#### 5.3.1 実験条件

- アルゴリズム : CLINK
- クラスタ数 : 5
- 最大検索文書数 : 1000
- 対象言語 : 日本語

まず、34 の検索課題ごとに、1 度のクラスタリングにおいて作成されたクラスタが適合文書を含むかどうかを調べる。次に、適合文書が 1 つのクラスタにしか出現しない課題について、再びクラスタリングを行う。以上の作業を適合文書が 1 つのクラスタにしか出現しない検索課題が無くなるまで繰り返す。

#### 5.3.2 結果

表 7 に結果を示す。これによると CLINK では、5 回目のクラスタリングの後初めて、クラスタサイズが元の文書集合サイズの  $1/2$  よりも小さくなっている。HBC の場合では、これが 1 回のクラスタリングで達成される<sup>6</sup>ことを考えると、文書サイズとクラスタリングの回数という観点だけからは、CLINK は HBC よりも劣ると言える。

## 6 考察

実験 1, 実験 2 では、最大検索文書数 (一度にクラスタリングする文書数) が少なければ、クラスタリングがより有効になると考えられる。また、クラスタのサイズに関するパラメータの設定の違いが評価の結果に大きな影響を与えることが判明した。

実験 2 では、5 段階の適合についての評価において、適合/不適合の 2 段階の場合の評価尺度を適用した。しかし 2 段階の場合と比べて結果に顕著な違いは見られなかった。これは、1 つのクラスタにおける 適合文書/不

<sup>6</sup>同条件で HBC の場合、1 回のクラスタリング後の適合文書を含むクラスタの平均文書数は 207.3、平均適合文書数は 7.362 であった。

適合文書が低い場合と比べ、クラスタのエントロピーがあまり変わらないことが原因の一つである。

また、クラスタリングアルゴリズムによって、作成されるクラスタのサイズは大きく異なる。実験1では最も良い結果であったCLINKが、実験3では、HBCと比べると極端に悪い結果となっている。これは、「クラスタのサイズを抑えることよりも適合文書をなるべく1つのクラスタに集中させる」と「適合文書を1つのクラスタに集中させることよりもクラスタのサイズを小さくする」ことが、トレード・オフの関係にあることを示している。

以上より、本研究で用いたタイプのクラスタリングの性能を単一の評価基準で測定することは難しく、複数の観点からの評価が必要であることがうかがえる。さらに、もし均等なサイズのクラスタを作るアルゴリズムのみの評価であれば、クラスタサイズを考慮する必要は少なくなるため、上述の観点が一つ減ることになり、評価が容易になると期待される。

## 7 おわりに

本稿では、クラスタリングの評価にエントロピーの概念を用い、検索結果のクラスタリングの妥当性を定量的に計測することを試み、既存の適合/不適合の判定に加えて、適合レベルを考慮した評価を行い、さらにクラスタリングによって作成されるクラスタのサイズとクラスタ回数に関する実験を行った。

今後は、クラスタサイズが均等になるように修正したアルゴリズムでの評価や、新聞、論文などの特許以外のテストコレクションを用いたクラスタリングの評価についても検討してゆきたい。

## 参考文献

- [1] D.R.Cutting, D.R.Karger, J.O.Pedersen, J.W.Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.318-329. 1992.
- [2] Atsushi Fujii and Tetsuya Ishikawa. Evaluating multilingual information retrieval and clustering at ULIS. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

- [3] 福井雅敏, 樋口重人, 藤井敦, 石川徹也. 日米対応特許コーパスを用いた対訳抽出手法. 情報処理学会研究報告 2001-NL-145, pp.23-28, Sep. 2001
- [4] Marti A Hearst. The Use of Categories and Clusters for Organizing Retrieval Results, In *Natural Language Information Retrieval*(T.Strzalkowski ed.), pp.333-374. Kluwer Academic Publishers. 2000.
- [5] Marti A Hearst and Jan O Pederson. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pp. 76-84, 1996.
- [6] Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. PRIME: A System for Multi-lingual Patent Retrieval. *Proceedings of MT Summit VIII*, pp.163-167, 2001
- [7] 樋口重人, 牧田光晴, 藤井敦, 石川徹也. 多言語検索システム PRIME. 言語処理学会第8回年次大会発表論文集, pp.196-199, 2002.
- [8] 岩山真, 藤井敦, 高野明彦, 神門典子. 特許コーパスを用いた検索タスクの提案. 情報処理学会研究報告 2001-FI-63, pp.49-56, 2001.
- [9] 岩山真, 徳永健伸. 確率的クラスタリングを用いた文書連想検索. 自然言語処理, 5(1), pp.101-117, 1998.
- [10] 牧田光晴, 樋口重人, 藤井敦, 石川徹也. 多言語検索システムの開発と評価. 情報科学技術フォーラム (FIT), 2002.(掲載予定)
- [11] 高野, 丹羽, 西岡, 岩山, 久光, 今一, 櫻井, 徳永, 奥村, 望月, 野本. 汎用連想計算エンジンの開発と大規模文書分析への応用. 情報処理振興事業協会「独創的情報技術育成事業」第19回 IPA 技術発表会, 2000.