# 大規模テストコレクションNTCIR-1とNTCIR-2 – レコードから見た違い –

栗山和子　吉岡真治　神門典子
国立情報学研究所 (NII)
{kuriyama,yoshioka,kando}@nii.ac.jp

**概要.** 本稿では、大規模テストコレクションNTCIR-1とNTCIR-2および、NTCIRワークショップ2について紹介し、NTCIR-1とNTCIR-2の文書レコードから見た違いについて述べる。具体的には、レコードの長さ、その分布、異なり語数、その頻度などから、NTCIR-1と2の違いについて考察する。

# Large Scale Test Collections NTCIR-1 and NTCIR2 – Differences in the Document Records –

Kazuko Kuriyama　Masaharu Yoshioka　Noriko Kando
National Institute of Informatics (NII)

**Abstract.** In this paper we describe large scale test collections, NTCIR-1 and NTCIR-2, and the 2nd NTCIR Workshop. We also show differences in the records of the test collections from point of view of statistical data; length of the records, their distribution, number of different terms, their frequencies, and so on.

## 1　はじめに

### 1.1　NTCIRプロジェクト

著者らは、国立情報学研究所（旧 学術情報センター）の「情報検索システム評価用テストコレクション構築」プロジェクトにおいて、情報検索システム評価用テストコレクション NTCIR（エンティサイル：NII-NACSIS Test Collection for Information Retrieval systems）の構築を行っている [12]。その過程において、1998年11月から1999年9月まで、テストコレクション1 (NTCIR-1)（予備版）を用いた評価型ワークショップとしてNTCIRワークショップ1を開催した[5],[6],[8],[9],[10],[11]。構築したテストコレクション1(NTCIR-1)は、1999年11月から研究目的での使用に限って公開している。また、2000年5月から2001年2月までは、テストコレクション2 (NTCIR-2)（予備版）を用いたNTCIRワークショップ2を開催し、NTCIR-2を構築中である。

本稿では、NTCIR-1、NTCIR-2、およびNTCIRワークショップ2について紹介し、NTCIR-1とNTCIR-2の文書レコードから見た違いについて述べる。具体的には、レコードの長さ、その分布、異なり語数、その頻度などから、NTCIR-1と2の違いについて考察する。

### 1.2　テストコレクションについて

テストコレクションとは、情報検索システムの検索性能評価に用いられる実験用セットのことであり、(1) 文書データベース、(2) 検索課題群、(3) 各検索課題に対する正解文書の網羅的リスト、からなる。

大規模テストコレクションとしては、TREC(Text REtrieval Conference)[3],[13] が有名である。TRECの登場以降、情報検索システムの評価実験は大規模テストコレクションを使用したものが主流となり、大規

模なテストコレクションを使った実験を行わなければ、情報検索の国際的なコミュニティに受け入れられにくいというのが現状である。そのようが状況から、日本でも情報処理学会データベースシステム研究会のワーキンググループにより、日本語の新聞記事を対象としたBMIR-J1,J2[7]が構築されたが、文書の種類と数量を増やすという意味から、NTCIRプロジェクトが開始された。

# 2 NTCIRワークショップ2

## 2.1 タスク

NTCIRワークショップの各タスクの概要は以下の通りである。詳しくは、http://www.rd.nacsis.ac.jp/~ntcadm/を参照されたい。

**中国語検索タスク (Chinese IR Tasks)**:中国語の単言語検索、および英語・中国語の言語横断検索。中国語の文書群に対して、新しい英語または中国語の検索課題で検索を行い、その検索有効性を調べる。使用するテストコレクションとしては、Chinese Information Retrieval Benchmark 1 (CIRB-1)を使用する。これは、(1)文書データベース(台湾の通信社5社から提供されたニュース記事132,173件)、(2)検索課題50件(英語・中国語各50件ずつ)、(3)正解文書リストからなる。詳しいタスクの説明については、http://www.lips.lis.ntu.edu.tw/cirb/index.htmを参照。

**日本語・英語検索タスク (Japanese and English IR Tasks)**: 日本語の単言語検索、英語の単言語検索、および英語・日本語の言語横断検索。訓練用セットとしては、テストコレクション1(NTCIR-1)の(1)文書データベース(学会発表論文の著者抄録約33万件、日本語・英語)、(2)検索課題83件(日本語のみ)、(3)正解文書リストを使用する。評価用セットとしては、テストコレクション2(NTCIR-2)(予備版)の(1)文書データベース(科学研究費補助金研究成果報告の要旨約30万件と学会発表論文の著者抄録約10万件、日本語・英語)、(2)検索課題49件(日本語、英語対訳で各49件ずつ)を使用する。

**自動要約タスク (Automatic Text Summarization Task)**: 毎日新聞の記事を用いて、日本語文書の自動要約を行う。目的は、日本語テキストに対する要約データの蓄積と、自動要約システムの評価である。詳しいタスクの説明については、http://galaga.jaist.ac.jp:8000/tsc/ を参照。

## 2.2 日程

2000年9月以降の主な日程は以下の通りである。

**2000年9月18日:** 日本語・英語検索タスクの検索結果提出
**2000年9月30日:** 中国語検索タスクの検索結果提出
**2000年9月:** 自動要約タスクのドライラン
**2000年11月:** 自動要約タスクの評価
**2001年1月10日:** 中国語検索タスク、日本語・英語検索タスクの正解判定結果の通知
**2001年2月1日:** 成果報告会用仮論文の原稿提出(全タスク)
**2001年2月21-23日:** 成果報告会

# 3 NTCIR-1とNTCIR-2

## 3.1 文書データ

### 3.1.1 NTCIR-1の文書データ

文書データベースは、国立情報学研究所(旧 学術情報センター)が提供している「学会発表データベース」の一部を使用している。「学会発表データベース」は、日本国内の65学協会の協力を得て、全国大会、研究会などの発表論文要旨(著者抄録)を集めたもので、NTCIR-1では、このデータベース中から約33万件を選定し、一般的に検索に用いられるフィールドである「標題」、「著者」、「発表学会」、「発表年月日」、「著者抄録」、「著者キーワード」を抽出し、文書データベースとして用いている。文書の約半数が日英対訳であり、文書データベースには、JEコレクション(日英文書全体)、Eコレクション(英語の標題と抄録を持つレコードの英語部分のみ)、Jコレクション(日本語の標題と抄録を持つレコードの日本語部分のみ)の3つのコレクションがある。文書データの例を図1に示す。

図 1. NTCIR-1 の文書データの例
（学会発表 DB、J コレクション）

```
<REC>
<ACCN>gakkai-0000011144</ACCN>
<TITL TYPE="kanji"> 電子原稿・電子出版・電子図書館-「SGML 実験誌」の作成実験を通して </TITL>
<AUPK TYPE="kanji"> 根岸 正光 </AUPK>
<CONF TYPE="kanji"> 研究発表会 (情報学基礎)</CONF>
<CNFD>1991. 11. 19</CNFD>
<ABST TYPE="kanji"><ABST.P> 電子出版というキーワードを中心に、文献の執筆、編集、印刷、流通の過程の電子化について、その現状を整理して今後の動向を検討する。とくに、電子出版に関する国際規格である SGML(Standard Generalized Markup Language) に対するわが国での動きに注目し、学術情報センターにおける「SGML 実験誌」およびその全文 CD-ROM 版の作成実験を通じて得られた知見を報告する。また電子図書館について、その諸形態を展望する。出版文化に依拠するこの種の社会システムの場合、技術的な問題というのは、その技術の社会的な受容・浸透の問題であり、この観点から標準化の重要性を論じる。</ABST.P></ABST>
<KYWD TYPE="kanji"> 電子出版 // 電子図書館 // 電子原稿 // SGML // 学術情報センター // 全文データベース
</KYWD>
<SOCN TYPE="kanji"> 情報処理学会 </SOCN>
</REC>
```

### 3.1.2 NTCIR-2 の文書データ

文書データベースは、国立情報学研究所が提供している「学会発表データベース」の一部と「科学研究費補助金研究成果報告概要データベース」の一部を使用している。

「学会発表データベース」からの文書データは、NTCIR-1 に収録後に新たに追加されたデータの中から約 10 万件を選定し、NTCIR-1 と同じフィールドを抽出したものである。

「科学研究費補助金研究成果報告概要データベース」からは、このデータベース中から約 30 万件を選定し、「報告年度」、「研究分野コード」、「研究課題名（標題）」、「研究報告概要（抄録）」、「図表の見出し」、「著者キーワード」を抽出し、文書データベースとして用いている。

学会発表 DB からの文書の約半数と科研報告 DB からの文書の約 25%が日英対訳であるが、文書データベースとしては、完全に日本語文書データと英語文書データを分け、英語文書の文書番号 (ACCN) を付けかえ、文書番号による対応付けはできないようにしてある。

コレクションとしては、NTCIR-1 の E コレクションと J コレクションに対応するものとして、E コレクションと J コレクションの 2 つのコレクションがあり、J コレクション（日本語文書データ）については、日本語解析ソフトウェア Happiness/BASE3.5[4] を用いて、テキストを自動的に語と語構成要素に分割した語分割テキストも提供している。語の区切（強い区切）は全角空白、語構成要素の区切（弱い区切）は全角アンダースコアである。

J、E コレクションについては NTCIR-1 とほぼ同じであるので、語分割テキストの例を図 2 に示す。

図 2. NTCIR-2 の文書データの例
（学会発表 DB、語分割テキスト）

```
<REC>
<ACCN>gakkai-j-0000441590</ACCN>
<TITL TYPE="kanji"> 大規模__テストコレクション__NTCIR - 1 の 構築 ( 1 ) - プーリング と 正解__判定 の 分析 - </TITL>
<AUPK TYPE="kanji"> 栗山 和子 / 江口 浩二 / 野末 俊比古 / 神門 典子 </AUPK>
<CONF> 全国__大会 </CONF>
<CNFD>1999. 09. 28 - 1999. 09. 30</CNFD>
<ABST TYPE="kanji"><ABST.P> 本__研究 の 目的 は 、 ( 1 ) 大規模__テストコレクション を 構築 する 手法 と して の プーリング の 有効性 を 検証 し 、 ( 2 ) プーリング__件数 が 検索__システム の 評価 に 関連 が ある か どうか 調べ 、 ( 3 ) 正解__判定 の 際 の 判定 の ゆれ が システム の 評価 に 関係 して くる か どうか を 明らか に する こと で ある 。</ABST.P><ABST.P>
(以下省略)
```

## 3.2 検索課題

検索課題は、分野の専門家（大学院生以上）から、インタビューあるいは一定の形式で形式の検索課題記入フォームによって収集された [5]。正解文書数が少なすぎて検索性能評価に影響を与えないように、予備検索を行い、正解文書数が 5 件以上あるものを選択している。

NTCIR ワークショップ 1 用には、検索システムの訓練用検索課題 30 件 (topic0001-0030) と評価用検索課題 53 件 (topic0031-0083) を用意し、両方の検索課題を合わせた 83 件を最終的なテストコレクション 1(NTCIR-1) の CD-ROM に納めた。NTCIR ワークショップ 2 では、検索システムの訓練用としては、NTCIR-1 の検索課題 83 件を使用し、評価用検索課題として新たに日本語検索課題と英語検索課題それぞれ 49 件を用意した。日英の検索課題は対訳になっている。

topic0001-0030、topic0031-0083、NTCIR ワークショップ 2 の評価用検索課題は、それぞれ概念（キー

ワード）の部分で若干形式が異なっている。図3にNTCIRワークショップ2の評価用検索課題の例を示す。

検索課題の形式は、初期のTRECの検索課題に準じ、SGML形式に類似したタグが付与されている（タグの詳細は、[5]を参照）。

図3. NTCIRワークショップ2の評価用検索課題の例

```
〈TOPIC q=0101〉

〈TITLE〉
B型肝炎
〈/TITLE〉

〈DESCRIPTION〉
遺伝子工学的手法によるB型肝炎ワクチンの開発について論じている文献
〈/DESCRIPTION〉

〈NARRATIVE〉
肝炎などのウイルス性疾患に対する安全かつ有効な予防法の確立は21世紀に向けての医療分野での重要な課題である。そのため、遺伝子工学的手法によるB型肝炎ワクチンの開発について論じていれば検索要求を満たす。開発されたB型肝炎ワクチンの物理化学的特性を論じているものやその免疫力増強に有効な免疫アジュバントについて論じているものも検索要求を満たす。しかし、遺伝子工学的手法に触れていない論文は不可。また、B型肝炎以外のワクチンも不可。
〈/NARRATIVE〉

〈CONCEPT〉
a. B型肝炎,
b. 遺伝子工学的手法,
c. ワクチン, 予防接種
〈/CONCEPT〉

〈FIELD〉
7. 医学・歯学
〈/FIELD〉

〈/TOPIC〉
```

### 3.3 正解文書リスト

大規模テストコレクション構築における正解文書候補の収集の手法としては、TRECなどで用いられているプーリング(Pooling)法がある。プーリング法では、同一課題に対して複数の検索手法を用いた複数の検索システムによる検索結果の上位一定数を集めてプールし正解文書候補として、それに対して正解判定を行う。

NTCIR-1の検索課題83件については、プーリング法を用いて、正解文書リストを作成した。正解文書候補収集の基本的な方法は、ワークショップ参加者から提出された検索結果の上位一定数のプーリングとし、そのプールに対して正解判定を行った。また、正解文書数が50件以上である検索課題については、対話型検索システムを用いたマニュアルでの追加検索を行い、正解文書候補を収集して、追加判定を行った。正解判定は、1つの検索課題について判定者2人で行い、最終的な判定は両者の協議に基づいて主判定者が行っている。判定は、「正解(A)」、「部分的正解(B)」、「不正解(C)」の三段階である。

以前の著者らの論文[8],[9],[10]では、プーリングの方法による正解文書リストの網羅性・公平性、複数判定者による正解判定の一致度の影響などについて検証し、評価用テストコレクションとして有効性に問題がないと考えられたので、NTCIRワークショップの評価用検索課題のための正解文書リストの作成についても、同様の方法で行う予定である。

## 4 統計的データ

### 4.1 レコードの長さ

NTCIR-1とNTCIR-2のJコレクション、Eコレクションのそれぞれのレコード数を表1に示す。NTCIRワークショップではNTCIR-1および2を使用する。

表1. レコード数

| 文書 | ntc1 | ntc2g | ntc2k | 計 |
|---|---|---|---|---|
| 日本語(J) | 332,918 | 116,177 | 287,071 | 736,166 |
| 英語(E) | 187,080 | 77,433 | 57,545 | 322,058 |

ntc1: NTCIR-1（学会発表DB）、ntc2g: NTCIR-2（学会発表DB）、ntc2k: NTCIR-2（科研報告DB）

NTCIR-1と2の日本語文書レコードの長さについてのデータを表2,3,4に示す。ntc1-j1はNTCI1R-1のJコレクション、ntc2-j0gはNTCIR-2のJコレクションのうち「学会発表データベース」から、ntc2-j0kは「科学研究費補助金研究成果報告概要データベース」から、それぞれ選定された部分である。

表2. ntc1-j1の長さ（学会発表DBから）（全332,918レコード）

| 文字数・個数 | TITL | ABST | | KYWD | 全体 |
|---|---|---|---|---|---|
| | 文字数 | 個数 | 文字数 | 個数 | 文字数 |
| max | 320 | 30 | 8,118 | 23 | 8,318 |
| min | 2 | 1 | 6 | 3 | 84 |
| ave | 28.5 | 1.3 | 284.8 | 4.1 | 439.6 |
| var | 135.6 | 0.6 | 16,852.8 | 2.6 | 205,918.8 |
| sdev | 11.6 | 0.8 | 129.8 | 1.6 | 453.8 |

TITL: 標題、ABST: 抄録、KYWD: キーワード、
max: 最長、min: 最短、ave: 平均、var: 分散、sdev:標準偏差

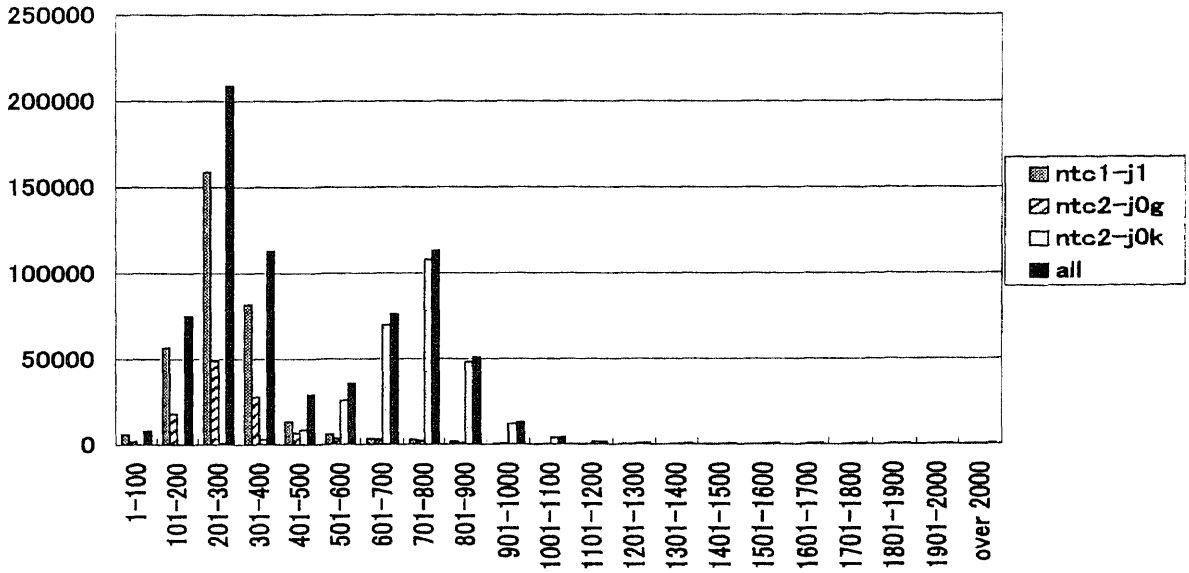


図 4. 抄録 (ABST) の長さの分布

表 3. ntc2-j0g の長さ（学会発表 DB から）
（全 116,177 レコード）

| 文字数・個数 | TITL | ABST | | KYWD | 全体 |
|---|---|---|---|---|---|
| | 文字数 | 個数 | 文字数 | 個数 | 文字数 |
| max | 169 | 31 | 4,327 | 16 | 4,480 |
| min | 3 | 1 | 4 | 3 | 83 |
| ave | 30.1 | 1.5 | 328.2 | 4.1 | 484.8 |
| var | 144.7 | 1.6 | 49,713.9 | 2.4 | 279,486.8 |
| sdev | 12.0 | 1.3 | 223.0 | 1.6 | 528.7 |

TITL: 標題、ABST: 抄録、KYWD: キーワード、
max: 最長、min: 最短、ave: 平均、var: 分散、sdev:標準偏差

表 4. ntc2-j0k の長さ（科研報告 DB から）
（全 287,071 レコード）

| 文字数・個数 | TITL | ABST | | KYWD | 全体 |
|---|---|---|---|---|---|
| | 文字数 | 個数 | 文字数 | 個数 | 文字数 |
| max | 148 | 44 | 8,851 | 29 | 9,018 |
| min | 1 | 1 | 4 | 1 | 54 |
| ave | 27.8 | 3.7 | 737.0 | 5.9 | 853.5 |
| var | 73.7 | 6.7 | 26,952.5 | 3.4 | 728,513.8 |
| sdev | 8.6 | 2.6 | 164.2 | 1.8 | 853.5 |

TITL: 標題（研究課題名）、ABST: 抄録（研究報告概要）、
KYWD: キーワード、max: 最長、min: 最短、ave: 平均、var: 分散、
sdev:標準偏差

また、図 4 に、ntc1-j1、ntc2-j0g、ntc2-j0k のそれぞれ、および、その和 (NTCIR-1 と 2 の日本語文書) (all) の抄録の長さ（文字数）の分布をグラフとして示す。

表 2,3,4、図 4 からわかるように、NTCIR-1 と 2 の日本語文書レコードでは、抄録の長さが異なっている。ntc2-j0g は、ntc1-j1 と同じ「学会発表データベース」から抽出された文書レコードである（ntc1-j1 作成後の追加分である）ので、最長 (max)、最短 (min) のレコードの主なフィールド中の長さに違いはあるものの、全レコードの平均 (ave) では、主なフィールドや全体の長さはほぼ同じであり、図 4 や標準偏差 (sdev) からわかるように、長さの分布も似ている。

それに対して、ntc2-j0k は、「科学研究費補助金研究成果報告概要データベース」という異なるデータベースから選定されたものであるので、標題（研究課題名）の長さは ntc1-j1 と同程度であるが、抄録（研究報告概要）の長さは平均で約 2 倍であり、図 4 の通り、分布全体が 2 倍の位置にずれていることがわかる。また、標準偏差から見て、抄録の分布は ntc1-j1 とよりは少し狭い範囲になっている。

NTCIR ワークショップ 2 では、NTCIR-1 と 2 の文書の両方を検索対象とするが、図 4 からわかるように、NTCIR-1+NTCIR-2 は、長さの分布の異なる文書集合の和であるので、分布が集中するところが二ヶ所ある。このように、長さの分布のピークが複数ある文書データベースが検索にどのような影響を与えるのかは、NTCIR ワークショップ 2 の検索結果の評価後、分析できれば、と考えている。

レコード全体の長さでは、ntc1-j1 と ntc2-j0g に対して ntc2-j0k は、レコードに含まれているフィールドが異なるので、表 2,3,4 の標準偏差からわかるように、

ntc2-j0k の方がばらつきがある。

## 4.2 異なり語数

NTCIR ワークショップ 2 では、海外からの参加者に対するサポートとして、あるいは、NTCIR-1 よりも詳細なシステム間の比較を可能にするために、NTCIR-1 と NTCIR-2 に日本語文書データと日本語検索課題について、あらかじめ語と語構成要素に分割したテキスト（語分割テキスト）を用意し、配布している。分割には日本語解析ソフトウェア Happiness/BASE3.5[4] を用い、テキストを自動的に分割している。語分割テキスト中では、語の区切（強い区切）は全角空白、語構成要素の区切（弱い区切）は全角アンダースコアとして分割を示している。

本節では、NTCIR-1 と NTCIR-2 の日本語文書データ (ntc1-j1, ntc2-j0g, ntc2-j0k) に対する語分割テキストデータ (ntc1-j1-wakachi, ntc2-w0g, ntc2-w0k) を使用し、異なり語数とその出現頻度を数えた。また、ntc2 として ntc2-w0g と ntc2-w0k の日本語文書データを合わせたもの、ntc-all として ntc1-j1-wakachi, ntc2-w0g, ntc2-w0k を合わせたものについても異なり語数と出現頻度を数えた。英語文書データについては、英文は空白で単語ごとに区切られており、異なり語を数える際には、複数の単語からなる複合語をどのように分割して 1 語 (term) とするかという問題があるので、本稿では扱わないことにする。

異なり語は、(1) 強い区切だけで区切った場合（語のみ）と (2) 強い区切と弱い区切だけで区切った場合（語と語校正要素）のそれぞれについて数えた。ただし、全角（2 バイト文字）のアルファベット、数字、ピリオド「.」、コンマ「,」は対応する半角（1 バイト文字）のアルファベット、数字、ピリオド「.」、コンマ「,」と同じものとみなし、半角の空白も強い区切とみなした。

表 5 に、ntc1-j1-wakachi、ntc2-w0g、ntc2-w0k のそれぞれについて、(1) 強い区切で切った場合と (2) 強い区切と弱い区切で切った場合、のそれぞれののべ語数、異なり語数、その出現頻度の平均、分散、標準偏差、変動係数を示す。表 6 に、ntc2(“ntc2-w0g”+”ntc2-w0k”) と ntc-all(“ntc1-j1-wakachi”+ “ntc2-w0g”+”ntc2-w0k”) のそれぞれについて、表 5 と同じ種類の値を示す。また、図 5,6 に、出現頻度の分布をグラフとして示す。ただし、単位は出現頻度の常用対数である。

表 5. 異なり語数 (ntc1-j1-w, ntc2-w0g, ntc2-w0k)

| segment | | ntc1-j1-wakachi | ntc2-w0g | ntc2-w0k |
|---|---|---|---|---|
| strong | total | 46,595,139 | 18,510,315 | 97,753,509 |
| | term | 2,265,248 | 1,114,708 | 3,929,792 |
| | ave | 20.6 | 16.6 | 24.9 |
| | var | 14,923,747.8 | 4,570,773.8 | 38,960,337.8 |
| | sdev | 3,863.1 | 2,137.9 | 6,241.8 |
| | cv | 187.8 | 128.7 | 250.9 |
| strong & weak | total | 56,044,058 | 22,038,575 | 114,180,102 |
| | term | 704,964 | 410,196 | 1,294,166 |
| | ave | 79.5 | 53.8 | 88.2 |
| | var | 48,842,723.7 | 12,609,845.4 | 119,663,096.6 |
| | sdev | 6,988.8 | 3,551.0 | 10,939.1 |
| | cv | 87.9 | 66.0 | 124.0 |

total: のべ語数、term: 異なり語数、ave: 出現頻度の平均、
var: 出現頻度の分散、sdev: 出現頻度の標準偏差、cv: 出現頻度の変動係数
strong: 強い区切、strong & weak: 強い区切と弱い区切

表 6. 異なり語数 (ntc2, ntc-all)

| segment | | ntc2 | ntc-all |
|---|---|---|---|
| strong | total | 116,263,824 | 162,858,963 |
| | term | 4,634,935 | 6,147,134 |
| | ave | 25.1 | 26.5 |
| | var | 46,117,160.4 | 67,849,169.4 |
| | sdev | 6,791.0 | 8,237.1 |
| | cv | 270.7 | 310.9 |
| strong & weak | total | 136,218,677 | 192,262,735 |
| | term | 1,482,052 | 1,831,047 |
| | ave | 91.9 | 105.0 |
| | var | 145,844,261.8 | 230,448,927.1 |
| | sdev | 12,076.6 | 15,180.5 |
| | cv | 131.4 | 144.6 |

total: のべ語数、term: 異なり語数、ave: 出現頻度の平均、
var: 出現頻度の分散、sdev: 出現頻度の標準偏差、cv: 出現頻度の変動係数
strong: 強い区切、strong & weak: 強い区切と弱い区切

前節で見たように、ntc2-j0k（科研報告 DB から抽出）の抄録は、平均では、ntc1-j1 と ntc2-j0g（学会発表 DB から抽出）の抄録の 2 倍の長さであった。表 5 からわかるように、異なり語数もほぼ 2 倍近くになっているが、一般的には、抄録の長さが長くなるのに比例して異なり語数が単調に増加するとは考えられない。しかし、同じ語が出現する回数は、抄録が長くなれば増えると考えられる。表 5 では、平均出現頻度 (ave) が増えており、出現頻度の標準偏差 (sdev) も増えていることから、1 つの語の出現回数は増えているが、少数回した出現しない語の数も増えているので、全体として抄録が長くなるほど、語の出現頻度はよりばらつくことがわかる。

表 6 から、ntc2,ntc-all とも、異なり語数は、表 5 のそれぞれデータベースでの異なり語の総和よりは小

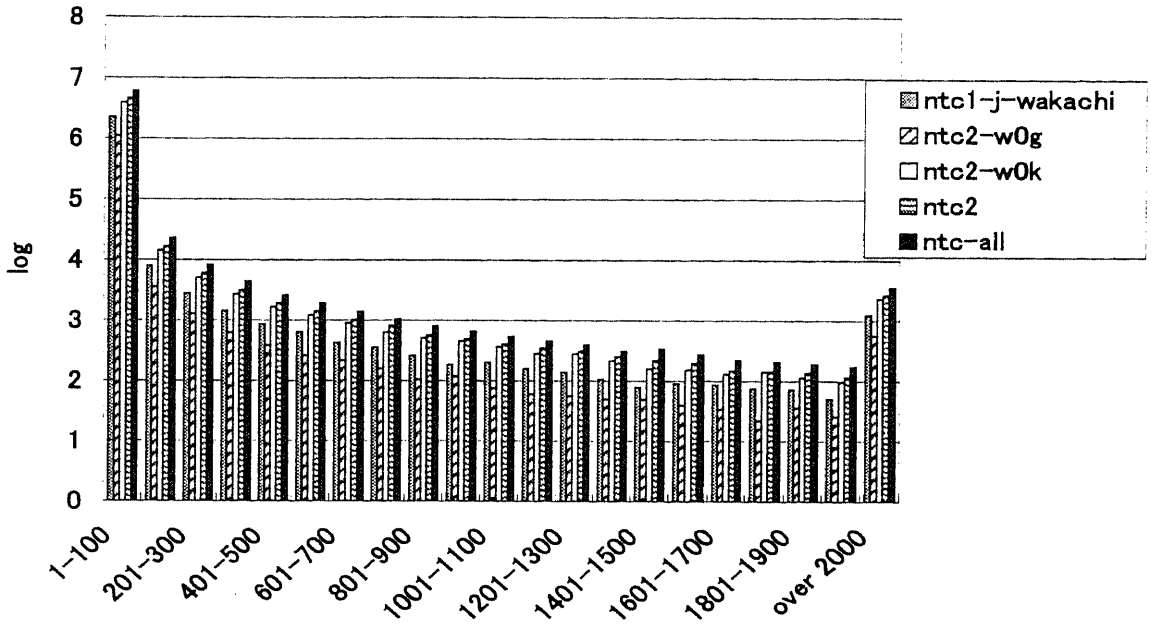


図 5. 語の出現頻度の分布（強い区切）

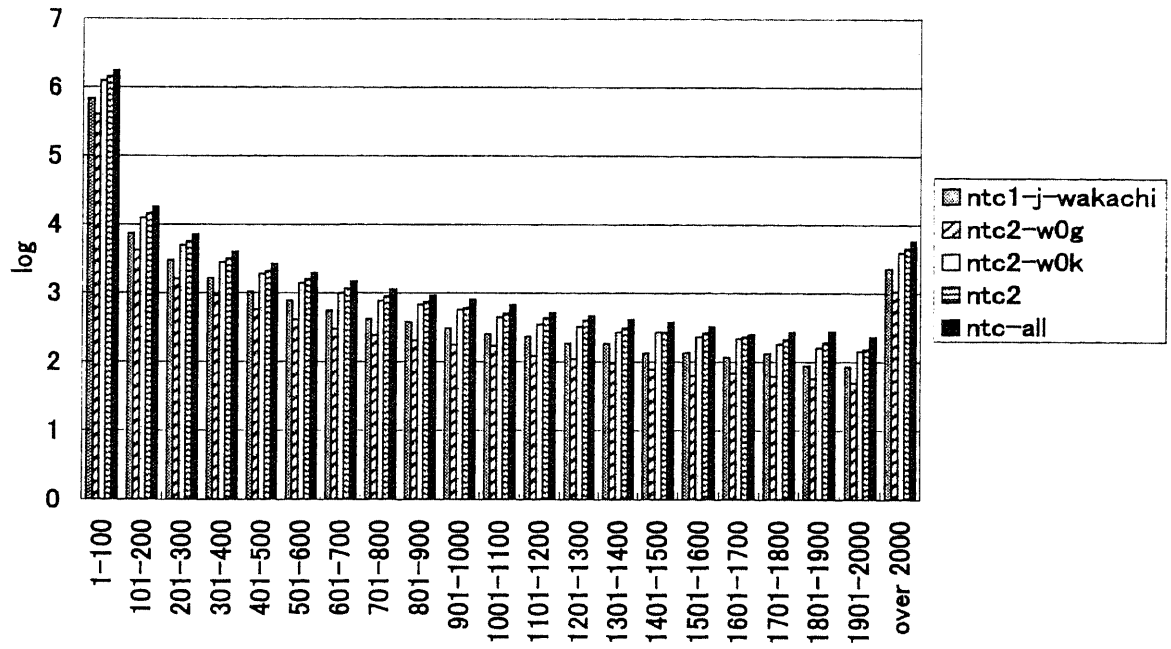


図 6. 語の出現頻度の分布（強い区切と弱い区切）

さくはなっているものの、複数のデータベースに重複して出てくる語がそれほど多くはないことがわかる。また、出現頻度の標準偏差からわかるように、異なるデータベースを合わせると、語の出現頻度はよりばらついてくる。

図5,6から、強い区切と、強い区切と弱い区切での語の出現頻度の傾向は、ほぼ同じようになることがわかる。

## 5 まとめ

情報検索において、索引語を自動的に識別する指標、ベクトル空間型モデルでの文書と検索質問との類似度、文書のカテゴリ分けなどでは、文書内に対する語の出現頻度のヴァリエーション（文書内での相対出現頻度と全文書における相対出現頻度との差や比、tfidf（「ある文書$d_i$における語$t_j$の出現頻度(Term Frequency)」×「逆文書頻度(Inverse Document Frequency)（$1/n_j$のヴァリエーション、$n_j$は語$t_j$が出現する文書数)」)、相澤が文献[1],[2]で述べているtfkli（「tf」×「カルバックライブラー情報量（Kullback-Leibler information)」）などが使われている。語の特徴量や個々の文書の分類ということではなく、コーパスとしてのデータベース全体の性質を比較するものとして、主にレコードの長さと語の出現頻度について述べたが、今後、上のような文書ごとの指標の観点からも詳細な考察が必要である。

現時点では、まだ、NTCIRワークショップ2での検索結果提出とその評価は行われていないので、文書データベースの性質の違いが検索に与える影響について述べることはできないが、今後の課題として、4章で述べたような、文書の長さや異なり語数などから見た文書データベースの違いが検索システムの検索結果やその評価にどのような影響を与えるか検討したい。

## 謝辞

本研究は、日本学術振興会未来開拓学術研究推進事業「高度分散情報資源活用のためのユービキタス情報システム」（課題番号JSPS-RFTF96P00602）による。

NTCIRプロジェクトの遂行にあたり、日本語解析ソフトウェアHappiness/BASE3.5による語分割テキストの公開を快諾してくださった（株）平和情報センターに深く感謝いたします。

本研究を行うにあたり、国立情報学研究所の影浦峡助教授から統計的手法に関するご助言をいただきました。ここに心から感謝いたします。

## 参考文献

[1] 相澤彰子. “語と文書の凶器に基づく「特徴量」の定義と適用”. 2000-FI-57-4, pp.25-32, 2000.

[2] Aizawa, A. “The Feature Quantity: An Information Theoretic Perspective of Thidf-like Measures”. In Proc. of ACM-SIGIR2000, pp.104-111, Athens, Greece, 2000.

[3] Voorhees, E.M., Harman, D. “Overview of the Seventh Text REtrieval Conference (TREC-7)”, NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7), 1999.

[4] Happiness/BASE3.5 for UNIX/WindowsNT ユーザーズマニュアル, 第3版.（株）平和情報センター, 1998.

[5] 神門典子ほか. “NTCIR-1：情報検索システム評価用テストコレクション構築の方針と実際”. 99-FI-53-5, pp.33-40, 1999.

[6] 神門典子ほか. “大規模テストコレクション NTCIR-1の構築 (2):検索課題の分析”. 情報処理学会第59回全国大会, pp.3-107-3-108, 1999.

[7] 木谷強ほか. “日本語情報検索システム評価用テストコレクション BMIR-J2”. 98-DBS-114, pp.15-22, 1998.

[8] 栗山和子ほか. “大規模テストコレクション構築のためのプーリングについて：NTCIR-1の予備テストの分析”. 99-FI-54-4, pp.25-32, 1999.

[9] 栗山和子ほか. “大規模テストコレクション構築について：NTCIR-1の訓練用検索課題の分析”. 99-FI-55-6, pp.41-48, 1999.

[10] 栗山和子ほか. “大規模テストコレクション NTCIR-1の構築 (1):プーリングと正解判定のの分析”. 情報処理学会第59回全国大会, pp.3-105-3-106, 1999.

[11] NTCIR Workshop 1: Proceedings of the Fist NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Tokyo, Aug.30-Sep.1, 1999.

[12] NTCIR: NII-NACSIS Test Collectioon for Information Retrieval.
http://www.rd.nacsis.ac.jp/~ntcadm/.

[13] Text REtrieval Conference (TREC).
http://trec.nist.gov/.