研究論文

文書頻度と節長を利用した図書概要縮約方式

Sentence Extraction of Book Abstract based on Document Frequency and Clause Length

小峰 恒
東京電機大学大学院工学研究科
Hisashi KOMINE
Graduate School of Engineering, Tokyo Denki University
山田 剛一
東京電機大学工学部
Koichi YAMADA
School of Engineering, Tokyo Denki University
絹川 博之
東京電機大学大学院工学研究科
Hiroshi KINUKAWA
Graduate School of Engineering, Tokyo Denki University
中川 裕志
東京大学情報基盤センター
Hiroshi NAKAGAWA
Information Technology Center, The University of Tokyo

要旨

近年，携帯電話や PHS を用いての図書検索サービスのニーズが高まっている．ところが，従来の図書検索の結果である図書概要はパソコンなどの大画面での閲覧を前提としているため，携帯電話など表示画面の小さい端末では閲覧しにくいものとなっている．検索結果を携帯端末の一画面に表示させるためには，図書概要を短く縮約する方式の開発が必要である．図書概要の縮約方式として，文書頻度（df）法による単語の重み付け方式に節の長さを複合させた縮約方式を提案し，実験評価している．実験評価では df 法，tf 法，tf･idf 法について，節の長さとの有無に関し，10-Fold cross validation を用いて実験評価した．その結果，df 法と節長の組み合わせによる複合型概要縮約方式が最も良い結果が得られることを確認した.

ABSTRACT

In recent years, the demand of Web pages browsing by using mobile terminals i.e. cellular phone, PHS etc. is increasing. However it is difficult to browse Web pages by using mobile terminals, because Web pages are made for large size display of personal computers. In order to solve this problem, it is necessary to summarize Web pages. In this paper, we propose a new method of sentence extraction based on document frequency and clause length for summarizing book abstract. We compared our method with conventional sentence extraction methods based on TF-method and TFIDF-method. Experimental results show that our method improves both recall and precision.

[キーワード]

要約，文書頻度，節長，図書概要

[Keywords]

Sentence extraction，Document frequency，Clause length，Book abstract

1　まえがき

1.1　本研究の目的

近年，技術の発達に伴い，携帯電話や PHS など Web ページを閲覧できる携帯端末が増えてきている．それに伴い，大学の図書館などで用いられている図書検索サービスをパソコンからだけではなく，携帯端末から利用したいというニーズが高まっている．ところが，図書検索の結果表示に使用する概要はパソコンを用いての閲覧を前提に設計されており，200 字前後で書かれている．そのため，表示領域が小さい携帯端末では，ほとんどの場合，図書概要が一画面では表示できないか，もしくは表示文字を小さくする必要がある．また，携帯端末では，通信コストは通信データ量に比例する．このためコスト面からも，通常の概要を携帯端末にそのまま表示することは好ましくない．

そこで，本研究では，携帯端末に直接検索結果を送るのではなく，無駄な情報を省いて重要な個所のみを抽出し，小さい画面に表示するのに適切な長さに縮約する方式を提案する．

1.2　研究対象

研究の対象として，東京大学情報基盤センター[1]の図書概要データベースであるブックコンテンツ内の図書データを利用しており，その例を図 1 に示す．

図書データのうち，当該図書の概要である Description（以下，概要と呼ぶ）を縮約の対象とする．およそ 1000 件の図書データの調査では概要構成文字数は 42～486 字であり，平均 181 字である．概要構成文数はおよそ三文となっている．

最近の携帯電話では一画面に表示できる文字数が 50～100 字であり，このため携帯端末に表示しようとすると，ほとんどの概要は一画面には納まらない．

以上より，図書データの概要を携帯端末表示画面の制約に合わせて，縮約することを目標とする．

ID: NB9430909
Title: ニューロ・ファジィ・遺伝的アルゴリズム
Series: エレクトロニクス実践シリーズ
Author: 萩原／将文【著】
Publisher: 産業図書
Year: 1994
Description: 本書では、新しい技術として特に大きな注目を集めているニューラルネットワークとファジィ、それから遺伝的アルゴリズムを総合的に扱い、個々の技術の背景から基本と応用、さらにそれらの融合方法までをポイントをおさえて解説しています。高校生程度の方々でも理解できるように、解説はできる限り平易になるよう心掛けました。
Contents: ・1章 はじめに
・2章 ニューラルネットワークとは何だろう
・3章 ニューラルネットワークの実際
・4章 ファジィの原理を学ぼう
ISBN: 4782855397

図 1　東京大学のブックコンテンツにおける図書データ例

2　図書概要の特徴と縮約

2.1　重要文抽出における従来の方式

重要文抽出型の要約手法[4, 5]として以下のものがある．

(1) 文の出現位置を基に文を抽出する方式

文書は，ある種のスタイルを持って書かれている場合が多い．その規則性を用いて，重要だと思われる文を抽出するのがこの方式である．

例としては，Lead 法が挙げられる[6]．Lead 法は段落の先頭の文を重要視し，文を抽出する方式である．新聞など，先頭に大まかな内容が書かれるような文書に対して非常に有効である．現に，NTCIR-2 TSC 課題 A-1（重要文抽出）では多くのシステムに Lead 法が用いられており，報道記事の要約においては有効な手法であると言える[7]．また，野畑ら[8]は段落の先頭だけではなく，末尾も重要であるという考えから，先頭に加えて末尾に近い文も重要文として重み付けをしている．

(2) 単語の重み付けにより文を抽出する方式

文書中に使用される単語を重み付けし，重みの高い語を重要とみなす．次に重要な単語が使われている文は重要であるという考えに基づき，文を抽出する方式である．

従来の単語の重み付けの方式としては，単語の出現頻度に基づく tf 法や，単語の出現頻度と特定の文書のみに出現することに着目した tf･idf 法がよく利用されている[9]．Mori らは検索結果表示向け文章の要約に対し，tf･idf 法に情報利得比を組み合わせた手法を提案している[10]．また，tf 法と文書頻度に基づく df 法をあわせた反復度を用いて，単語の重み付けを行っている研究もある[11]．また，単語の重み付けによる方式である tf 法，tf･idf 法と Lead 法を組み合わせた手法もある[8, 12]．

(3) 特定の言語表現，特徴を含む文を抽出する方式

特定の分野の文書には，重要個所に用いられる手がかり表現が存在する場合がある．その情報を基に，

文を抽出するのがこの方式である．情報抽出では，よく用いられている方式である[13]．要約では，特許広報や講演文といった分野の文書に対してこの方式が用いられている[14, 15]．

手がかり表現の抽出・作成は，人手による方式と機械支援による方式[16]とがある．また，文長も文の特徴の一つである[8]．手がかり表現や文長などの文の特徴を Lead 法や tf･idf 法と組み合わせた方式もある[8]．

2.2 図書概要の例と特徴

図書概要の例を示す．

下線部は，概要の中で重要な箇所であり，縮約文となりうると人手で判断される個所を示している．

例 1

本書は，光・レーザをいろいろな工学分野に応用するという視点に立って書かれている．したがって，応用する際に必要となる基礎的内容だけを，できるだけ直感的に理解しやすいように説明した．

例 2

初心者にとって良き入門書であると同時に，熟達した研究者には実験手法の視野を広げ新しい発想のヒントになるように配慮した新しいタイプの実験書．多くの実験室で共通の基礎的手法であり，学部４年生あるいは修士課程の院生にとってその習得は必須．測定系の制御やデータ収録・解析に不可欠なエレクトロニクス，センサー，トランスデューサー，コンピュータの具体的かつ実用的な解説を行う．

図書概要の書き方には以下の特徴がある．

(1) 図書概要は本の内容を簡単に説明する必要があるために，短く書かれている場合がほとんどである．そのために使われる単語が少なく，例 1 のように，全ての単語の単語頻度が"1"であることも少なくない．このような文書では，tf 法では，単語の重み付けの際，差が明確に現れないため，抽出すべき個所を特定するのに有効であるとは言えない．

(2) 例 2 のように，図書の概要は新聞のように重要部が先頭に書かれているとは限らない．そのため，Lead 法を用いても重要文を抽出できるとは限らない．

(3) 『本書』という出だしで書かれた文中には当該図書の要点が簡潔に書かれていることが多い．例 1 のように「本書は～～書かれている」とある場合，この内容を見れば，その部分が図書の内容を書いてあるということは予想できる．

(4) 『解説』は，図書の内容をよく表す文には多く使われている．図書概要の例における網かけ部分がそれにあたる．このように，図書概要では重要箇所によく用いられる単語がいくつか見られた．

(5) 『本書』や『解説』などは多くの図書概要で使われており，それらを含む文には概要の要点が書かれていることが多く，文構成上から見ると，述語に直接係る文節の省略が少ない傾向が見られた．

(1)(2)より，図書の概要から重要文を抽出する方式として，**2.1** 節で述べた出現位置方式や単語頻度を重みとする方式は有効でないと考えられる．また，(3)(4)(5)より，重要文抽出方式として図書概要の文の特徴を利用する方式が有効であると考えられる．

図書概要の特徴である，本の内容を示す文によく使われる『本書』や『解説』などの単語（例では網かけ部分）を，文章特徴語と呼ぶことにする．文章特徴語は多くの図書概要中で使われており，文書頻度（df）値が高いと予想される．

2.3 文書頻度と文章特徴語

図書概要において，文書頻度が高い単語と文章特徴語との関連について，以下の図書により調べた．

① 『生命』＆『バイオ』
② 『建物』＆『建築』
③ 『物理』＆『運動』
④ 『エレクトロニクス』
⑤ 『自然』＆『環境』＆『研究』
⑥ 『電気』＆『通信』
⑦ 『現代』＆『経済』＆『金融』
⑧ 『ロボット』
⑨ 『政治』＆『倫理』

以上の，9 つの学術系のキーワード群によって検索された図書，計 3019 件を対象に概要に出現する各単語を，文書頻度の降順に並べ，その結果を表 1 に示す．割合は，3019 件の図書全体に対し，どれほどの出現率であるかを示す．文書頻度で並べる場合『が』や『の』のように助詞などが上位にのぼるが，それらの機能語や非自立語は文章特徴語としては不適切だと考えられるため，調査対象となる単語を名詞の自立語もしくは未知語のみとした．

表 1 から『本書』や『解説』など，図書の説明に

重要だと思われる単語は文書頻度（df）値が高い．しかし，文書頻度の高い単語の中には，『経済』や『現代』など図書概要の要点を表す文の手がかりには必ずしもならないが，分野共通の語なども含まれていた．分野共通語も図書内容の要点を示すのに重要であることと，文章特徴語のほとんどを含むことから，文章特徴語と分野共通語の集合を，高 df 値の語で近似する[3]ことにした．

2.4　抽出すべき概要単位

抽出する概要単位を調べるために，三つの図書データの概要を構成する文数と文構成文字数，及び本方式で用いる節構成文字数の平均を調査し，その結果を表 2 に示す．

図書概要の重要個所は一つの文ではなく，複数の文に含まれることがわかった．ところが，表 2 より文を二文抽出すると 100 字を超えることが多いため，携帯端末の表示には長すぎる．そのため，今回の抽出する概要単位は表層上の文ではなく，節にし，短くすることにする．節とは，句読点で区切られた単位と定義する．ただし，以下の条件を満たす場合，区切らないこととする．

(1) 直前が接続詞，係助詞の場合

接続詞は後述する節と意味的に一体であるため区切らない方が良く，係助詞の場合はその直前が主格になることが多いからである．

(2) 連続した名詞，未知語の区切りとして使われている読点

読点は列挙の区切りとして使われている場合があり，その内容は続いているからである．

2.5　正解節と概要全体の節長比較

概要全体とシステムが抽出すべき節（以降，正解節と呼ぶ）の節長比較のため，平均文字数，平均単語数を調査し，表 3 に示す．ここで言う単語とは，名詞，未知語を指す．その結果，正解節は全平均に比べて単語数が多い．これは図書の内容を示す節では，主要な文節の省略が少ないからであると考えられる．よって，正解節を抽出する方式として節長も重要な手がかりであるといえる．

3　図書概要縮約方式

3.1　図書概要縮約処理手順

2 章での調査，考察を基に，図書概要縮約方式として文書頻度に節長を組み合わせた方式を提案する．

処理手順は以下のとおりである．

(1) ブックコンテンツ・データベースから図書データの概要を抽出する．

(2) 抽出した概要に形態素解析を行い，品詞情報を得る．なお，本研究では形態素解析器として茶筌[2]を用いた．

(3) 品詞情報を用い，**2.4** 節の定義に基づいて，概要を節に分割する．概要の節数を p とする．

(4) 節ごとに算出対象語を抜き出す．算出対象語とは，文章特徴語や分野共通語の候補となりうる単語のことを指す．本方式では，品詞が名詞もしくは未知語の単語とする．ただし，『こと』や『もの』など，単体では意味を持たない形式

表 1　学術系図書概要における単語の文書頻度

順位	単語	文書頻度	割合（%）
1	本書	955	31.6
2	解説	475	15.7
3	技術	446	14.8
4	研究	398	13.2
5	問題	303	10.0
6	基礎	268	8.9
7	経済	268	8.9
8	現代	266	8.8
9	書	253	8.4
10	分野	244	8.1
11	環境	232	7.7
12	日本	228	7.6
13	者	223	7.4
14	社会	220	7.3
15	世界	209	6.9
16	情報	205	6.8
17	理解	196	6.5
18	科学	185	6.1
19	理論	180	6.0
20	応用	177	5.9

表 2　概要を構成する文

	概要の構成する文数	文構成文字数の平均	節構成文字数の平均
ｴﾚｸﾄﾛﾆｸｽ	2.90	51.06	27.2
物理	3.76	48.42	26.8
法律	2.73	58.60	27.0
全体	3.12	52.88	27.0

表 3　概要全体と正解節の節長の比較

分野	全体		正解節	
	文字数	単語数	文字数	単語数
ｴﾚｸﾄﾛﾆｸｽ	27.2	5.65	36.1	8.52
物理	26.8	6.16	37.3	9.21
法律	27.0	6.11	37.1	9.71

名詞は含めない．節 i(1 ≦ i ≦ p)内の算出対象語の個数（単語数）を k_i とし，節 i の j(1 ≦ j ≦ k_i)番目の算出対象語を w_{ij} と表す．算出対象語の数が多いほど，主要な文節の省略が少なく情報量が多い，重要な節とみなす．

(5) 抽出した算出対象語 w_{ij} の重みとして，df，tf，tf･idf 値を計算する．以降それぞれ df_{ij} ，tf_{ij} ，$tf \cdot idf_{ij}$ と表す．以下に重み付けの意味を述べる．

(a) df 値：文書集合，本稿では図書概要の集合の中で，多くの図書概要に使用されている特定表現の単語，例えば，文章特徴語や分野共通語などを重要とみなす．

(b) tf 値：個々の図書概要の中で出現頻度の高い語を重要とみなす．

(c) tf･idf 値：個々の図書概要中で出現頻度が高く，かつそれを含んでいる図書概要の頻度が低い語を重要とみなす．

(6) 節中の算出対象語の重みの和と節長を表す単語数から節の重みを計算する．節の重みの計算方法は **3.2** 節で述べる．なお本方式では，節長を単語数で表すことにした．

(7) (6)による節の重みの高い節を抽出する．なお，本実験では携帯端末表示可能な文への縮約を目標とするので選択する節は二節とした．

3.2 節の重み付け方式

(1) 節 i での，df 値, tf 値, tf･idf 値による算出対象語 w_{ij} の重みの和をそれぞれ $Wweight_{df}(i)$，$Wweight_{tf}(i)$，$Wweight_{tf \cdot idf}(i)$ とする．

(a) df 値による $Wweight_{df}(i)$

$$Wweight_{df}(i) = \sum_{j=1}^{k_i} df_{ij} \qquad (式1)$$

(b) tf 値による $Wweight_{tf}(i)$

$$Wweight_{tf}(i) = \sum_{j=1}^{k_i} tf_{ij} \qquad (式2)$$

(c) tf･idf 値による $Wweight_{tf \cdot idf}(i)$

$$Wweight_{tf \cdot idf}(i) = \sum_{j=1}^{k_i} tf \cdot idf_{ij} \qquad (式3)$$

ただし，(1 ≦ i ≦ p， 1 ≦ j ≦ k_i).

(2) 節長 $length(i)$ は節内の単語数 k_i とする．

(3) 各単語の重み，節長といった異なる重みを複合するため，各要素を以下の方法で正規化する．

df 値　：全文書数で割って正規化する．

tf 値　：各文書内の全単語数で割り正規化する．

tf･idf 値　：上記の二つの値を用いて正規化とする．

節長　：全文書における節単位での単語数の平均で割って正規化する．

(4) 単語の重み付けの際，df 値の場合は df 値が"1"，tf 値，tf･idf 値の場合では tf 値が"1"の単語は単語の重み計算時に含めないこととする．tf 値，tf･idf 値において，算出対象語に tf 値"1"の単語を含めると，節を構成する全ての単語の重みの和を取ることになる．これにより，節長と組み合わせたとき，節長の重みが二重でかかってしまうので本項の扱いとした．

(5) 節 i の重みは，(1)の各算出対象語の単語の重みの和と，(2)の節長とから，以下の式により算出する．

(a) df 法：単語の重み付けを df 値とし，節長と組み合わせた方式である．

$$Pweight_{df}(i) = (1-\alpha_{df})Wweight_{df}(i) + \alpha_{df} \cdot lenght(d) \qquad (式4)$$

α_{df} ：df 法の節長比重　(0≦ α_{df} ≦1)

(b) tf 法：単語の重み付けを tf 値とし，節長と組み合わせた方式である．

$$Pweight_{tf}(i) = (1-\alpha_{tf})Wweight_{tf}(i) + \alpha_{i} \cdot length(i) \qquad (式5)$$

α_{tf} ：tf 法の節長比重　(0≦ α_{tf} ≦1)

(c) tf･idf 法：単語の重み付けを tf･idf 値とし，節長と組み合わせた方式である．

$$Pweight_{tf \times idf}(i) = (1-\alpha_{tf \cdot idf})Wweight_{i} + \alpha_{tf \cdot idf} \cdot lenght(i) \qquad (式6)$$

$\alpha_{tf \cdot idf}$ ：tf･idf 法の節長比重(0≦ $\alpha_{tf \cdot idf}$ ≦1)

(6) (5)によって算出された節の重み $Pweight_{df}(i)$，$Pweight_{tf}(i)$，$Pweight_{tf \cdot idf}(i)$ のそれぞれ上位二節を概要として選択する．

4 図書概要縮約方式の評価実験

単語重みと節長を組み合わせた図書概要縮約方式を評価実験する．具体的には，単語重み付け方式として，本稿で提案する df 法と，tf 法，tf･idf 法に関し，節長の組み合わせを使用しない場合（以下，単語重み単独と呼ぶ）と使用する場合（以下，節長複合と呼ぶ）について，10-Fold cross validation により比較実験する．節長複合の場合，各 α_{df} ，α_{tf} ，$\alpha_{tf \cdot idf}$ の値も同様に 10-Fold cross validation によって得た値を用いる．

4.1 実験対象データ

エレクトロニクス，物理，法律の三つの分野の図書データを

- エレクトロニクス：『エレクトロニクス』
- 物理：『物理』＆『出版年 1998 年～2001 年』
- 法律：『法律』＆『国』

各分野『』で括られた検索項目によって検索した図書データを各分野のデータとする．ただし，抽出する節数を二と定めたので三つ以上の節で構成される図書データのみを対象とする．

実験対象図書データについて，分野ごとに図書数と，概要構成節数の平均を調査し，その結果を表 4 に示す．

4.2 正解データ作成方法

以下の手順で正解データを作成した．

(1) 三人の正解データ作成者に節で区切られた概要を示し，図書の特徴を示す節を二つ選択させる．
(2) (1)の結果について，正解データ作成者の多数決により，上位二つを正解の節とする．

本実験では上記により作成した正解節を，システムが抽出すべき節とする．

4.3 実験の流れ

3.2 節(式 4) (式 5) (式 6)の節長比重 α_{df} ， α_{tf} ， $\alpha_{tf\cdot idf}$ に関して，節長複合の場合は 10-Fold cross validation の学習データを用いて，それぞれの値を 0.005 刻みで取り，精度，再現率が最も高くなる α_{df} ， α_{tf} ， $\alpha_{tf\cdot idf}$ を算出し，テストの際にその値を用いて評価を行う．

単語重み単独の場合は，節長を用いないので α_{df} ， α_{tf} ， $\alpha_{tf\cdot idf}$ の値は 0 で，学習は必要ない．そのため，以下の (3)(4)及び(6)の手順を含まない．

節長複合の場合の実験の流れを以下に示す．

(1) 図書集合を無作為に 10 等分する．
(2) 分割した一つをテストデータとし，残りを学習データとする．
(3) 学習データから名詞，未知語を抽出し，df 値，tf 値，tf･idf 値と節長を算出し，**3.2**(3)で示した方法により正規化を行う．
(4) (3)で計算した単語の重みと節長を複合する．節中の単語の重みの和を $Wweight_{df}(i)$ ， $Wweight_{tf}(i)$ ， $Wweight_{tf\cdot idf}(i)$ ，節長を $length(i)$ として節の重み $Pweight_{df}(i)$ ， $Pweight_{tf}(i)$ ， $Pweight_{tf\cdot idf}(i)$ を計算する．α_{df} ， α_{tf} ， $\alpha_{tf\cdot idf}$ の値を 0.005 刻みで変化させ，値（df 値，tf 値，tf･idf 値）ごとに精度，再現率が最も高い値を算出する．算出した α_{df} ， α_{tf} ， $\alpha_{tf\cdot idf}$ が，各単語の重みと節長を複合する際に最も適切であるとする．
(5) テストデータに図書概要縮約方式を用いる．図書概要から名詞，未知語を抽出し，df 値，tf 値，tf･idf 値と節長を算出する．計算の際，df 値，idf 値は学習データでの値を，tf 値はテストデータでの値を用いる．
(6) (5)で算出した各重み付けを **3.2**(3)で示した方法により正規化を行い，単語の重みと節長を複合する．節中の単語の重みの各和を $Wweight_{df}(i)$ ， $Wweight_{tf}(i)$ ， $Wweight_{tf\cdot idf}(i)$ ，節長を $length(i)$ として，(4)にて算出した各単語の重みに最適な α_{df} ， α_{tf} ， $\alpha_{tf\cdot idf}$ を用いて節の重み $Pweight_{df}(i)$ ， $Pweight_{tf}(i)$ ， $Pweight_{tf\cdot idf}(i)$ を計算する．
(7) それぞれの手法において，節の重みの上位二節を抽出し，精度，再現率を算出する．
(8) 実験データを変更し，残りの 9 つのデータを学習データとして(3)へと戻る．
ただし，変更されるデータは，まだテストデータとして用いていないものとする．分割されたデータすべてがテストデータとして用いられたのなら，(9)へと進む．
(9) (7)によって 10 回算出された精度，再現率の平均を求める．

4.4 実験結果

正解の節数は **4.2** 節に記した通り，一文書につき二節となるので，『文書数×2』節になる．また，抽出する節数も，一文書につき二つと定めたので，同様に『文書数×2』節となる．つまり，

$$精度＝再現率＝\frac{抽出した正解節数}{文書数\times 2} \qquad (式\ 7)$$

となり，本実験では精度と再現率は等しい．

単語重み単独および節長複合の df 法，tf 法，tf･idf 法による図書概要縮約方式について，評価実験し，それぞれの精度，再現率を求め，その結果を表 5 に示す．太文字は各分野における最大値を示す．各単

表 4 実験対象図書データ

分野	図書数	節数（平均）
エレクトロニクス	314	5.46
物理	310	6.02
法律	339	6.78

語の重み付けによる，節の抽出例を**付録 1** に示す．**付録 1** 表内の『正解』とは，正解作成時にその節が重要であると判断した人数である．

4.5 実験結果の考察

(1) 得られた精度，再現率

① 表 5 より，単語重み単独実験では実験に用いた三つの実験データすべてにおいて，df 法がもっとも性能が高く，tf 法が次，tf･idf 法が一番低いという結果が得られた．

② 節長複合実験では，節長を複合することにより，df 法に関しては 2~10%程度，精度，再現率が上がることがわかった．

③ 単語重み単独実験及び節長複合実験において df 法と，tf 法，tf･idf 法とを比較した結果を表 6 に示す．図書概要の縮約に関して df 法の精度，再現率は，単語重み単独の場合 2~8%，節長複合の場合 1~4%，tf 法，tf･idf 法より高い．

(a) 以上より，図書概要の縮約に関して，df 法に基づく重要節抽出方式の有効性を確認できた．また，節長を複合することにより，いずれの単語重み法も精度，再現率が上がり，df 値と節長の複合方式が最も高い結果となった．

(i) df 値の高い語は，文章特徴語や分野共通語など多くの図書概要に使用されている特定表現の単語であり，概要文をさらに縮約する場合，有効な手がかりになることを示している．

(ii) tf 値は，200 字前後の図書概要では値の差が小さく，重要節抽出にそれほど有効でないことを示している．

(iii) tf･idf 値は，特定の図書概要においてよく使われている語に着目するが，(ii)より tf 値の値の差がさほど大きくないことに加えて，図書概要が短いことからそれぞれの分野の中で個々の図書を特徴づける語の出現も少ないと考えられる．また，(i)で挙げた，多くの図書概要に含まれる特定表現語の tf･idf 値は小さく，tf･idf 法では特定表現語を含まれない節を重要とみなしてしまい，一番低い結果になったと考えられる．

(b) 本実験では，平均 6 節程度（表 4）の文章から二節抽出しており，要約率は 33%程度となっている．NTCIR-2 TSC 課題 A-1（重要文抽出）では，要約する文の対象が新聞記事ではあるが，要約率 30%では最大値は 0.518，要約率 50%で F 値の最大値は 0.633 となっている[7]．本実験で得た精度，再現率は **4.4** 節の(式 7)と F 値の定義から，F 値と言える．単純に比較は出来ないが，表 5 の結果は，評価できる値であると考える．

(c) df 値の高い語を手がかり語としてみなす本方式は，従来の方式（**2.1** 節(3)）で用いられる機械学習による手がかり語表現の収集に比べて，簡便であるが故に高速な収集を可能としている．

(d) 本稿では，図書概要について実験評価した．図書集合と異なる文書集合に対しても，図書概要縮約方式の有効性を調べる必要がある．

(2) 作成した縮約文の読みやすさ

節抽出により作成した縮約文は，以下に示す問題点があり，滑らかで読みやすい日本文に直すことが必要となる．例として，単語重み単独及び，節長複合方式による，図書概要の節抽出による要約結果の一部を**付録 1** に示す．

表 6　df 法と tf 法，tf･idf 法の差

	df 法と tf 法の差		df 法と tf･idf 法の差	
	単独	複合	単独	複合
エレクトロニクス	0.020	0.007	0.040	0.038
物理	0.077	0.021	0.077	0.039
法律	0.029	0.008	0.059	0.032

表 7　作成した縮約文の文字数の平均

	df 法	df 法 + 節長	正解
エレクトロニクス	71.7	80.8	72.2
物理	70.9	83.7	74.6
法律	71.2	82.0	74.2

表 5　単語重み単独実験，節長複合実験の精度，再現率

分野	df 法			tf 法			tf･idf 法		
	#1:単独	#2:**複合**	#2-#1	#3:単独	#4:複合	#4-#3	#5:単独	#6:複合	#6-#5
エレクトロニクス	0.624	**0.681**	0.057	0.604	0.674	0.070	0.584	0.643	0.059
物理	0.562	**0.669**	0.107	0.485	0.640	0.155	0.485	0.630	0.145
法律	0.611	**0.628**	0.017	0.582	0.620	0.038	0.552	0.596	0.044

(a) 節の区切りが文の途中となった場合に，文として成り立たない．抽出例(3)，(5)では，文の途中で節が区切られている．文の前半だけ抽出しているため，文章として完結しておらず，文章の意味が通じない．この解決のため，後述の未抽出の節中から動詞を補う，もしくは，語尾の動詞を終止形に変形させるなどの処理が必要となる．

(b) 抽出した節の結束性が問題となる場合がある．抽出された節にある指示語や代名詞で指し示されている語句が，縮約文として抽出されない場合，縮約文中に先行詞が無いため，内容を理解できない．これは重要文抽出において一般に言えることである．抽出例(6)では，指示語はないが，言葉が省略されており，その単語が抽出されないため，抽出文だけでは，違う意味に取られてしまう．指示語や代名詞が指し示す先行詞を未抽出の節から探し出し，置き換えることが必要となる．

(c) 抽出した節には，冗長に書かれているために無駄な単語が存在する場合がある．また，異なる文の節を抽出した場合に，単語が重複する場合もある．**付録 1** の抽出例(1)，(2)，(5)など，重複している単語や，無駄な単語などがあり，それらの単語を削除し，読みやすさを向上させる必要がある．

(d) df 法による単語重み単独方式によって作成された縮約文，節長複合方式によって作成された縮約文，人手により作成された正解節による縮約文，以上の三通りの縮約文の平均文字数を表 7 に示す．すべてにおいて，平均文字数が 70～80 文字程度となっている．携帯端末の機種によっては，一画面に表示できる文字が 50 字程度であり，表示方法の工夫が必要である．

5 むすび

5.1 成果のまとめ

(1) 本論文では，図書概要の縮約に関して，図書の内容をよく表す箇所に用いられる単語（文章特徴語や分野共通語など）と節長に着目し，文書頻度(df 値)に節長を複合した図書概要縮約方式 df 法を提案した．

(2) df 法を，単語重みをそれぞれ tf 値，tf･idf 値とした tf 法，tf･idf 法と比較した．

(a) 10-Fold cross validation による評価の結果，単語重み単独の精度，再現率は df 法が 56.2~62.4%となり，tf 法，tf･idf 法より 2~8%高く，三つの集合すべてにおいて，df 法がもっとも性能が高かった．

(b) 単語重みに節長を含めることによって df 法で 2~10%程度，精度，再現率が上がった．

(c) 節長複合による df 法の精度，再現率が 62.8~68.1%となり,tf 法，tf･idf 法との比較の結果,df 法が他手法より 1~4%高く，df 法と節長の複合方式が最も高い結果を得られた．

5.2 今後の課題

概要縮約方式の今後の課題は以下のとおりである．

(1) 読みやすい縮約文の生成

本方式では，節を抽出したに過ぎない．そのために改善すべき問題がある．

(a) 抽出単位を節にしたことによる文章の不整合．

(b) 抽出した節の結束性の欠如．

(c) 重複，あるいは無駄な単語の存在．

以上の問題点を改善して，読みやすく文章として適切な縮約文を作成可能な処理方式の検討が必要である．また，表示領域の小さい機種に対して，表示方法の検討も必要である．

(2) 他の文書集合に対する適用可能性の検討

提案方式の図書概要以外への適応可能性の検討が必要である．

謝辞

実験データとして東京大学情報基盤センターのブックコンテンツ，形態素解析に奈良先端科学技術大学院大学の茶筌を使用させて頂きました．関係者の方々に深く感謝致します．また，正解データの作成に御協力頂いた著者所属の計算言語学研究室の方々にお礼申し上げます．

参考文献

[1] 東京大学情報基盤センター，「ブックコンテンツ」 http://contents.lib.u-tokyo.ac.jp/contents/top.html.

[2] 奈良先端科学技術大学院大学，「形態素解析器茶筌」，http://chasen.aist-nara.ac.jp/.

[3] 小峰恒; 絹川博之; 中川裕志，「単語の文書頻度と文の長さを利用した抄録縮約方式」，『情報処理学会自然言語処理研究報告』，NL-149, p.73-80, 2002.

[4] 奥村学; 難波英嗣，「テキスト自動要約に関す

る研究動向」, 『自然言語処理』, Vol.6, No.6, p.1-26, 1999.

[5] Mani, I., "Automatic Summarization", Amsterdam, John Benjamins Publishing Company, 2001, 285p.

[6] Edmundson, H. P., "New methods in automatic abstracting.", *Journal of ACM*, 16(2), p.264-285, 1969.

[7] 難波英嗣; 奥村学, 「第２回ＮＴＣＩＲワークショップ自動要約タスク（ＴＳＣ）の結果および評価法の分析」, 『情報処理学会自然言語処理研究報告』, NL-144, p.143-150, 2001.

[8] 野畑周; 関根聡; 井佐原均; Ralph, Grishman, 「自動獲得した言語的パタンを用いた重要文抽出システム」, 『言語処理学会第８回年次大会』, p.539-542, 2002.

[9] Salton, G.; Yang, C. S., "On the Specification of Term Values in Automatic Indexing.", *Journal of Documentation*, 29(4), p.351-372, 1973.

[10] Mori, T.; Kikuchi, M.; Yoshida, K., "Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems", *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, p.5-205 - 5-212, 2001.

[11] 武田善行; 梅村恭司, 「キーワード抽出を実現する文書頻度分析」, 『情報処理学会自然言語処理研究報告』, NL-146, p.27-32, 2001.

[12] Ishikawa, K.; Ando, S.; Okumura, A., "Hybrid Text Summarization Method based on the TF Method and Lead Method", *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, p.5-219 - 5-224, 2001.

[13] 桝井文人; 福本淳一, 「製品情報一覧の自動提示のための情報抽出」, 『言語処理学会第６回年次大会ワークショップ』, p.56-63, 2000.

[14] 原正巳; 木谷強; 江里口善生, 「特徴的表現を利用した特許概要作成法の検討」, 『情報処理学会自然言語処理研究報告』, NL-100, p.105-112, 1994.

[15] 伊藤山彦; 松本賢司; 谷口泰郎; 柏岡秀紀; 田中英輝, 「講演文を対象にした重要文抽出」, 『言語処理学会第７回年次大会』, p.305-308, 2001.

[16] Sudo, K.; Sekine, S.; Grishman, R., "Automatic pattern acquisition for Japanese information extraction", *In Proceedings of Human Language Technology Conference*, 2001.

[17] 徳永健伸,「言語と計算５　情報検索と言語処理」,辻井潤一, 東京, 東京大学出版会,1999, 234p.

付録１．単語重み単独方式及び，節長複合方式による図書概要の節の抽出例

以下に凡例を示す.

単語重み単独　：各単語に，df 法，tf 法，tf･idf 法により重み付けし，節の重みを算出する節抽出方式.
節長複合　　：各単語に，df 法，tf 法，tf･idf 法により重み付けしたものに，節長の重みを複合させて，節の重みを算出する節抽出方式.
正解　　：正解データ作成者のうち、当該節が正解節であると判断した人数．"2"以上を正解節とする.
本文　　：節単位に分割された図書概要．本文の各行を連接すると，図書概要の本文全体となる.
☑　　：各方式が選択した節を示す．各方式による節の重み付けの大きい順から二節選択する.

抽出例（１）

節長複合			単語重み単独			正解	本文
df	tf	tfidf	df	tf	tfidf		
☑	☑	☑	☐	☐	☐	2	超LSIという半導体の小片に秘められた大きな力は、コンピュータや通信をはじめとする産業界全体の変革を引き起こしてきた。
☐	☑	☑	☐	☑	☑	1	天文学的な数の世界と電子顕微鏡的な超微細の世界がとじこめられている超LSI。
☐	☐	☐	☑	☐	☐	0	まさに現代技術の粋を集めた産物だ。
☑	☐	☐	☑	☑	☑	3	本書は、超LSIの技術の特質とわたしたちの生活に与えるインパクトの大きさを明らかにする。

抽出例（２）

節長複合			単語重み単独			正解	本文
df	tf	tfidf	df	tf	tfidf		
☑	☐	☐	☑	☐	☐	3	本書では、新しい技術として特に大きな注目を集めているニューラルネットワークとファジィ、それから遺伝的アルゴルズムを総合的に扱い、
☑	☑	☑	☑	☑	☑	3	個々の技術の背景から基本と応用、さらにそれらの融合方法までをポイントをおさえて解説しています。
☐	☐	☐	☐	☐	☐	0	高校生程度の方々でも理解できるように、
☐	☐	☐	☐	☐	☐	0	解説はできる限り平易になるよう心掛けました。
☐	☐	☐	☐	☐	☐	0	同時に、
☐	☐	☐	☐	☐	☐	0	現場の技術者の方々の製品開発にも十分役に立つように、
☐	☑	☑	☐	☑	☑	0	技術の流れと根本的な原理、応用における注意点やポイントを解説しました。

抽出例（３）

節長複合			単語重み単独			正解	本文
df	tf	tfidf	df	tf	tfidf		
☑	☐	☐	☑	☐	☐	2	本書では、メゾスコピック系のさまざまな量子伝導現象を紹介し、
☐	☐	☐	☐	☐	☐	0	その原理と応用について解説する。
☐	☐	☐	☐	☑	☑	1	拡散領域・バリスティック領域の現象はもとより、
☐	☐	☐	☐	☐	☐	0	特に、
☑	☑	☑	☐	☐	☐	3	極微細素子において注目されている単一電子トンネリング現象のもととなるクーロン・ブロッケイドについて詳説する。
☐	☑	☑	☑	☑	☑	0	応用科学と純粋科学が強く影響し合いながら発展してきたこの分野の、
☐	☐	☐	☐	☐	☐	0	基礎研究と応用の現状を把握するのに格好の入門書。

抽出例（４）

節長複合			単語重み単独			正解	本文
df	tf	tfidf	df	tf	tfidf		
☐	☐	☐	☐	☐	☐	0	空を飛ぶ鳥や昆虫の大きさは、飛翔のメカニズムとどのように関わりをもつのか、
☐	☐	☐	☐	☐	☐	0	野球やサッカーのボールはなぜ曲がるのか、
☑	☑	☑	☑	☑	☑	2	身のまわりの疑問を流体力学によって説明する。
☑	☐	☐	☑	☐	☐	3	本書では、生物、パラシュート、みそ汁等、特に身近で親しみやすい現象をとりあげる。
☐	☑	☑	☐	☑	☑	0	また、ボールまわりの空気の流れに関連して、
☐	☐	☐	☐	☐	☐	1	ラケットの力学にも触れる。

抽出例（５）

節長複合			単語重み単独			正解	本文
df	tf	tfidf	df	tf	tfidf		
☐	☐	☐	☐	☑	☑	0	この辞典は、現代の国語生活に立脚し、
☑	☑	☑	☑	☑	☑	3	現代語を中心に古語や百科語をも含めた総合的な国語辞典として編集したものである。
☐	☐	☐	☐	☐	☐	0	収録した語は、日常用いる現代語はもとより、
☑	☑	☑	☑	☐	☐	3	万葉集・源氏物語をはじめとするわが国古典にあらわれる語、医学・生物学・物理学・法律・経済など各専門分野における用語、および地名・人名・作品名などのいわゆる固有名詞など、
☐	☐	☐	☐	☐	☐	0	約22万語におよぶ。

抽出例（６）

節長複合			単語重み単独			正解	本文
df	tf	tfidf	df	tf	tfidf		
☑	☑	☑	☐	☐	☐	3	雇用主に対する罰則を含む不法就労対策、在留資格の整備、入国審査手続の簡易化など、
☐	☑	☑	☑	☑	☑	0	入管法の大幅な改正作業に参画した編者らが、
☐	☐	☐	☐	☑	☑	1	立案作業、各省庁との協議、国会審議を踏まえて、
☑	☐	☐	☑	☐	☐	2	改正の内容や運用上の問題点を分かり易く解説したＱ＆Ａ。