社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (20)
2008－ＮＬ－183 (20)
2008／1／22

科学技術白書の計量的分析による科学技術政策の可視化

石塚 隆男
亜細亜大学経営学部

本研究は、経時的な出現順序をもつ文書データの構造を明らかにするために文章中にある単語が初めて出現する位置情報をもとに新語数に関するいくつかの指標を提案する。対象データとして、過去20年にわたる『科学技術白書』の目次を用い、提案した指標によりわが国の科学技術政策を可視化することを試みた。指標の中では、IDF値によるウエイトを付加した自己情報量が他の指標よりも特徴を際立たせるのに有効であることが判明した。本研究で提案する手法は、前提知識や試行錯誤的な調整が不要なトップダウン的アプローチであり、全体構造を俯瞰するには有用であると考えられる。

Visualization of Policy on Science and Technology by Quantitative Analysis of White Paper on S&T

Takao Ishizuka
Faculty of Administration, Asia University

We propose a new visualization method of large text structure. Our method uses the number of new words in each document/paragraph. We quantize a self information content of each document by using the number of new words and a generalized inverse document frequency weights. The new index was applied to the series of the White Paper on Science and Technology over 23 years, and the extracted features coincided with the actual events or policies. Our approach uses no assumption nor heuristics, and is applicable to texts in every field.

1．はじめに

本研究は、あるテーマに関して刊行された年鑑や年次報告書のように、出現に時間順序がある大規模な文書データの系列を計量的に分析することにより文章構造を同定し、特徴を可視化する手法を検討することを目的とする。特徴の可視化は、一連の文章全体の要約にとって重要なプロセスであるが、本研究は要約作業そのものには立ち入らない。

このような逐次刊行物は、たとえば、企業の財務・決算報告書やCSR・環境報告書など枚挙にいとまがないが、本研究では長期にわたり政府により公刊されている『白書』を取り上げることにした。わが国には2005年時点で毎年40種類を超える白書が刊行されているが、その中でも時代の変化をとらえやすく、分析的意義があると考えられる『科学技術白書』を対象とする。科学技術白書は、政府のWebページによれば「科学技術の振興に関する年次報告」の通称であり、昭和33年版以降のものについて文部科学省のWebページから閲覧することができる。ただし、ファイル形式はテキスト形式とPDF形式が混在している。

科学技術白書を対象とした理由のひとつは、わが国の科学技術政策・予算の全貌並びに方向性が一般国民には見えにくい点にある。膨大な科学技術予算が、どこに使われているのかを経年的に把握することも一般国民には難しい。残念ながら、科学技術白書には予算の具体的な使途や金額は明記されていないので科学技術政策と予算情報と照合するためには別の情報源（たとえば、官報）にアクセスする必要があるが、限界がある。

わが国の科学技術政策を知るためには『科学技

術白書』によるのが妥当と思われるが、そのコンテンツや構造体系について論じた文献は見当たらない。内容構成には科学技術白書の編集責任者である文部科学省科学技術・学術政策局調査調整課長の見識や意向が反映されているといわれている。仮に個人的な意向により編集されたものだとしても、白書は省内の稟議を経て毎年の国会で承認され初めて世に出る公文書であり、分析の意義を損なうものではないと考える。

今回、1985年（昭和60年）～2007年までの毎年の科学技術白書の目次を対象に名詞の単語を中心とした分析を行った結果、いくつかの知見が得られたので報告する。

2．データの構造と可視化指標

本研究が対象とする文書データの構造と可視化のための指標について説明する。

一般にある程度まとまった長文の文章は、複数のパラグラフから構成されている。パラグラフは、著者の言いたいことのまとまりであり、文脈や文章内容を理解する上でパラグラフ情報を活用するのはきわめてナイーブなアプローチであると考えられる。形式的なパラグラフは、最初の１字が空白で始まる箇所を見つけることにより認識することができる。文章構造は、パラグラフの順序やパラグラフ間の関係によって構造を特徴づけることができる。パラグラフ間の関係を媒介しているのは、文章を構成する単語である。

Salton et al. (1996)は、テキストセグメントへの経時的分割とサブトピックへの意味的分割の相互作用によりテキスト構造の特徴づけがなされ、情報検索や要約にも活用できるとしている。

さて、本研究が対象とする逐次刊行物も同様に時間順序に意味があり、毎期の刊行物テキストをひとつのパラグラフとみなせば、対象全期間のテキスト全体がひとつの文章に相当する。

ベクトル空間モデルにしたがえば、文書データは構成する単語を要素とする高次元ベクトルとして記述できるが、本研究では構成する単語を要素とする集合により文書を表現する。

式(1)に示すように、対象とする文章の全体Tは、文書データT_j（j＝1～N）の連接により構成されているとする。

$$T = T_1 \mid T_2 \mid \cdots \mid T_N \qquad (1)$$

Tがひとつの文章であれば、T_jは第jパラグラフの文章に相当し、逐次刊行物であれば、第j期の刊行物文書に相当する。

文書T_jの要素数＝総単語数をn(T_j)と書くことにする。T_jの中には、重複して出現する単語もあるので重複を除いた総単語の種類数をv(T_j)と書くことにする。

文書はT_1, T_2, ・・・, T_Nの順で出現しているため、それ以前の文書にはなく、文書T_jにおいて新たに出現した単語＝新語数X_jを式(2)により逐次求めることができる。

$$\left.\begin{aligned} X_1 &= v(T_1) \\ X_2 &= v(T_2) - v(T_1) \\ X_3 &= v(T_3) - v(T_1 \cup T_2) \\ &\cdots\cdots \\ X_N &= v(T_N) - v\left(\bigcup_{j=1\sim N-1} T_j\right) \end{aligned}\right\} \qquad (2)$$

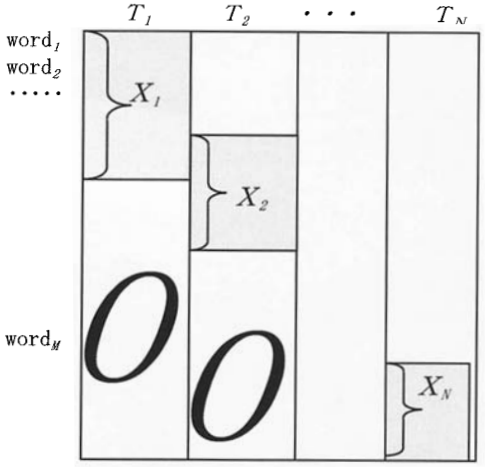

図１．単語×文書マトリクスの構造
（ は、新語を表す）

図１は、本研究の対象とする文章データの構造を示したものである。新語数の次元だけを見れば、各文書ベクトルは直交していることがわかる。

文章全体Tは、新語数の総和M＝ΣX_jの次元×構成する文書数Nの次元に展開されるが、一般に、構成するすべての単語や文書が同程度に重要であることはほとんどない。特徴抽出には、データの次元を代表的な少数の次元に縮約する作業が不可欠である。

本研究では、新語数に関する分布並びに統計的指標により文章の構造を集約する方法について検討を行う。本来、ある単語が“新語”かどうかは個々の読者にとっての話であるが、読者の知識レベルは無視し、上述のように当該文章内の出現位置により新語かどうかを判定することにする。

文章全体の中で相対的に新語の多い文書は、著者が新語により紙幅を費やした文書であり、新語が少ない文書よりも多くの情報を読者にもたらすと考えられる。新語を含まない文書の重要度が低いとは限らない。しかし、対象とする文章に関する予備的な情報や知識なしに、しかも効率的に情報を取捨選択し処理することが求められてい

る状況下において膨大な文書集合の中から新語の多い文書から優先的に読むのはきわめて妥当なアプローチと考えられる。

新語数に関する分布や統計的な指標を以下に提案する。

1）新語数の分布： $\{X_1, X_2, \cdots, X_N\}$

新語数の分布から相対的に新語が多い文書を抽出することが可能である。

2）新語率の分布：

式(3)に示すように各文書の新語数を当該文書の総単語数で除することにより新語率を計算し、分布を図示することができる。

$$\text{文書}T_j\text{の新語率} = X_j / n(T_j) \quad (3)$$

新語率は、0～1の値をとる。新語数と新語率のどちらがよいかは一概にはいえないが、一般的に文書のボリュームが大きくなるにつれ、新語数が増加するため、紙幅に関係なくスケールフリーで評価したい場合には新語率がよいだろう。

3）自己情報量

新語率をもとに式(4)により各文書の自己情報量(石塚(2005))を計算することができる。

文書T_jの自己情報量：

$$I_j = -\log(1 - X_j / n(T_j)) \quad (4)$$

$X_j = 0$のとき、$I_j = 0$となる。$X_j = n(T_j)$のとき、$I_j = \log(n(T_j)+1)$となる。式(4)の自己情報量は新語率を強調することができるが、総単語数$n(T_j)$の大きい文書の場合、新語率は小さな値となり、自己情報量も小さくなる。どの単語も同じウエイトで単純に単語数をカウントするため、どの文書にも出現する一般語の影響を受ける等の問題点がある。この点については、4）で改善案を説明する。

4）一般化 IDF ウエイト付き自己情報量

上述の自己情報量の問題点を解決するために、各単語に重要度に応じて何らかのウエイトを付加した上でカウントすることが考えられる。対象とする文書集合のすべてに出現する単語は一般語または全体のテーマに直接関係する語であり、新語としての重要度は低い。逆に特定の文書にのみ出現する単語はまさに新語であり、重要である。こうした目的に適したウエイトとして IDF（= Inverse Document Frequency）が知られている（Spärck Jones(1972)）。IDF は、直観的に理解しやすいだけでなく、多くのバリエーションが存在するが、近年の研究で情報量との関連等、理論的な解釈や裏づけが示されている（相澤(2000), Robertson(2004)）。

文章全体Tの中で単語 $word_i$ が出現した文書数をd_iとすれば、IDF 値は式(5)で計算される。

式(5)では、IDF 値をウエイトとして用いるため、［0，1］の範囲に基準化している。

単語$word_i$の IDF 値：

$$IDF_i = \frac{\log(N/d_i)}{\log N} = 1 - \frac{\log d_i}{\log N} \quad (5)$$

IDF 値は出現文書数の増加に伴い、急激に減少するため、目的に応じてウエイトを変化させたい場合には式(6)で表現される一般化 IDF（以下、GIDF と呼ぶ）値を用いることが考えられる。

$$GIDF_i = 1 - \left(\frac{\log d_i}{\log N}\right)^a \quad (6)$$

N＝50 で a の値を変化させたときの$d_i = 1$～Nに対する GIDF 値のグラフを図２に示す。

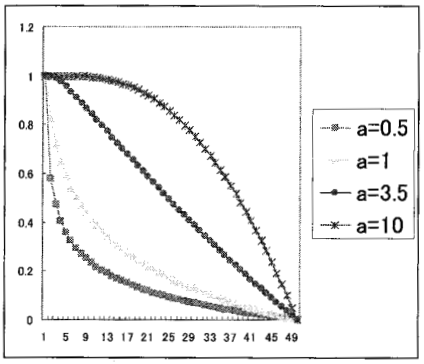

図２．パラメータ a に対する GIDF 値の変化

図２より a＝３～４でほぼ直線的に GIDF 値が変化し、a の値を十分大きくとると式(6)の第２項が０に近づくため等ウエイトになることがわかる。

GIDF ウエイトを用いた自己情報量の計算には新語数の代わりに、文書Ｔｊの新語について GIDF 値を合計したものを用いる。単語$word_i$の文書T_jにおける出現頻度をTF_{ij}とする。文書T_jの全単語について IDF 値を合計することは、$TF_{ij} \times GIDF_i$を$_i$について合計することに等しい。したがって、文書T_jの GIDF ウエイト付き自己情報量GWI_jは式(7)で計算される。

$$GWI_j = -\log\left(1 - \left[\frac{\sum_{i\in\text{新語の}} GIDF_i}{\sum_i TF_{ij} \cdot GIDF_i}\right]\right) \quad (7)$$

5）“瞬間”新語率

信頼性工学における故障率(真壁(1985))の考え方を用い、第ｊ番目の文書T_jにおける“瞬間”新語率λ_jを式(8)により計算することができる。

$$\lambda_j = \frac{X_j}{\sum_{k=1}^{N} X_k - \sum_{k=1}^{j-1} X_k} \quad (8)$$

瞬間新語率は、第ｊ番目以降に出現する文書中の新語数に対する文書T_j中の新語の割合を示している。故障率のアナロジーにしたがえば、λ_j

が一定か、減少型か、増加型により文章構造の類型化が可能になる。

6）文書のクラスター分析

以上述べた指標は1次元的な表現しかできないが、文章構造の可視化に有効な多変量解析手法としてクラスター分析や多次元尺度構成法がある。本研究では、各文書T_jについて構成する単語の出現湯現頻度を要素とするベクトルで表現し、階層的クラスター分析を行う。

3．関連研究

本研究の方法論的な目的はテキストデータの構造を可視化することであり、本章ではこれまでの先行研究や関連研究を概観し、本研究との異同について述べる。

テキストを意味的パラグラフ等の部分テキストに分割するテキストセグメンテーションに関して多くの研究がなされている。隣接する部分テキスト間の結びつきの強さを定量化するのが一般的な考え方である。Kozima(1993)は、テキストを単語の窓で切り取り、語彙結束性を測定し、その谷を検出することによりテキストの分割を行う方法を提案している。Hearst(1997)は、テキストをトークン系列化し、隣接するいくつかの系列によりブロックを構成し、ブロック間のコサイン類似度の変化を調べるTextTilingアルゴリズムを提案している。Nakao(2002)は、TextTilingを改良し、話題階層（＝サブトピック）を検出するアルゴリズムを提案している(Mani(2003))。望月他(1995)は、接続詞や文型などの表層的な情報も用い、意味的パラグラフへの分割を提案している。平尾他(2000)は、比較的短い文章に適用可能な方法として語彙的結束性とIDFによる単語の重要度を相補的に結合した方法を提案している。別所(2001)は、対象テキストと同分野のコーパスから単語の概念ベクトルを生成し、用いる方法を提案している。

以上のように、従来のテキストセグメンテーションのアプローチの多くは、語彙的結束性や語彙的連鎖を拠り所に窓の移動により境界を見つけるボトムアップ的手法が大半であり、窓の幅の決定がヒューリスティックであることやシソーラスやコーパス等の対象テキスト外の情報に依存し、汎用性の高い方法とはいえない。本研究はトップダウン型のアプローチであり、単語×文書マトリクス以外の知識を用いないため分野を問わず計量的分析を可能にしている。

テキストセグメンテーションに関する研究の動向は以上のとおりであるが、次にテキスト構造に関する近年の研究を概観する。Baeza-Yates et al(1996)は、内容と構造を統合した文書検索の必要性を述べ、6つのモデル定性的な評価検討を行っている。Hernandez et al(2003)は、検索された結果の個々のテキストが何について書かれているのかを効率的に判断する観点からグローバル並びにローカルのトピックを検出するシステムを提案しており、Nakao(前掲)がトピック階層を要約に活用しているのと同様の発想に基づいている。Bateman et al(2001)は、情報表現の自動化を検討する中で修辞構造理論（Rhetorical Structure Theory）を用い、テキストをツリー構造で表現し、分析を行っている。以上のことから、テキスト構造の利用目的により構造の表現方法は変わりうるとしてもトピックの階層構造には普遍性があり、トピック抽出は今後ますます重要な研究領域であると考えられる。

4．研究方法

対象とした文書データは、1985年～2007年までの毎年の科学技術白書の目次である。本文はテキスト形式とPDF形式が混在しており、全体像を把握するのに本文レベルの文章は必要ないと判断した。むしろ、重要な概念や用語は目次の章、節、項の見出しに含まれており、一般的な語彙が少ないため、分析の質の向上が期待できると考えた。なお、書籍版とWeb版で目次の粒度が異なる箇所があり、書籍版に統一した。

科学技術白書はこれまでのところ、

第1部　科学技術振興の成果
第2部　科学技術活動の状況／動向
第3部　科学技術の施策

の3部から構成され、今日に至っている。書籍版の大きさが平成12年度(2000)以前はA5判であったが、平成13年度(2001)からはA4判になり、カラー頁が増加した。

各年の科学技術白書の目次をテキスト化し、データベースを作成した。目次には本文目次の他に図表目次も付加されており、近年はグラフィカルな説明が多用されていることから図表目次もデータ化した。

各年の目次データを形態素解析し、単語×年別白書マトリクスの形に集計した。形態素解析は文字コードの種別を用い、ひらがなをストップワードとする独自プログラムによった。科学技術白書には複合語の出現頻度が高く、また、「もんじゅ」、「しんかい」、「かぐや」等のひらがなの固有名詞が登場するため、これらを抽出できるようプログラム化した。23年分の白書の本文目次から名詞を中心に約1400語の単語が抽出された。

5．結果

図3に年別部別ページ数の変化のグラフを示す。図4に年別部別構成割合の変化のグラフを示す。図5に年別の目次における本文項目数、図表項目数並びに新語数の変化のグラフを示す。

1994年の白書では、第1部で「今、世界の中で」と題し、主要国の科学技術動向の紹介を行い、過去最大のページ数となった。2001年版からA4判への変更に伴いページ数が減少し、ここ数年は300頁前後で推移している。図4より大きな政策

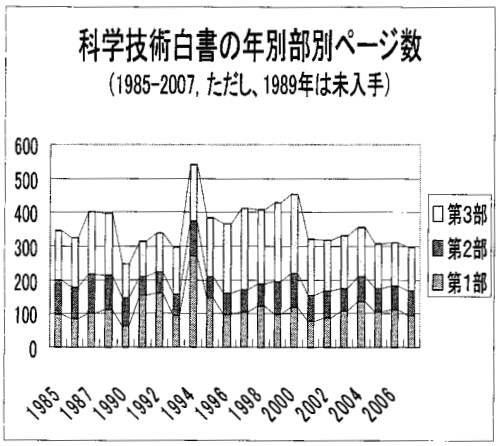

図３．科学技術白書の年別部別ページ数

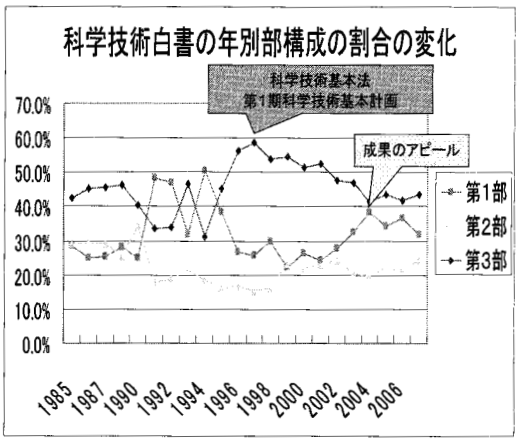

図４．科学技術白書の年別部構成の割合の変化

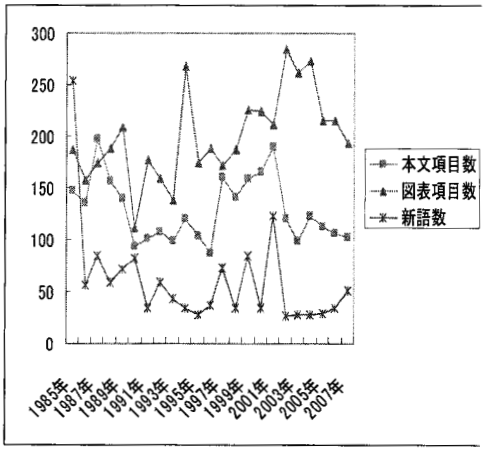

図５．科学技術白書の目次項目数と新語数の推移

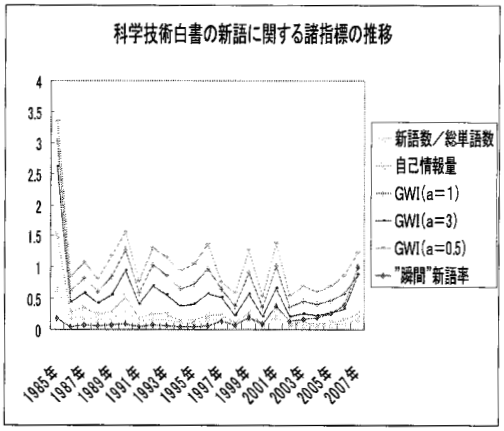

図６．科学技術白書の新語に関する諸指標の推移

の実施はページ数の構成割合に現れていることがわかる。図５より本文目次の項目数は 1996～2000 年の第２期科学技術基本計画において増大し、ここ数年は本文、図表項目とも減少傾向にあることがわかる。

図６は、新語に関する諸指標の経年的変化をグラフ化したものである。各指標とも 1990 年、92 年、96 年、99 年、2001 年に極大値を示している。図５からこれらの年の白書は新語数が相対的に大きいことがわかる。2001 年は科学技術庁から文部科学省に移行した年であり、制度の変更が如実に出たものの単年度のみの変化にとどまり、2002 年度にはその影響は出ていない。

GIDF のウエイトを用いた自己情報量 (GWI) は、べき乗値のパラメータ a の値が小さいほど少数の高ウエイトと多数の低ウエイトになるため、a＝0.5 のグラフが最も高い値を示している。単純な自己情報量よりも IDF ウエイトを付加した方が特徴が強調されることがわかる。

“瞬間” 新語率は、99 年と 01 年に極大値を示しているが、GWI ほど感度は高くないことがわかる。

今回の “瞬間” 新語率の計算には、新語数を用いたが、IDF を用いれば改善される可能性がある。

図７は、出現時期に特徴のある単語を示したものである。かつては使用されていた単語が 2001 年を境に使用されなくなる一方、新たに使用され始めた単語があることがわかる。

図８は、ウォード法によるクラスター分析を行った結果のデンドログラムである。科学技術政策は毎年頻繁に変わるものではなく、何ヵ年計画というように慣性力が強いため、隣接年の白書間にクラスターが形成されていることがわかる。第１期科学技術基本計画 (1996-2000 年)、第２期 (2001－2005 年)、第３期 (2006-2010 年) にほ

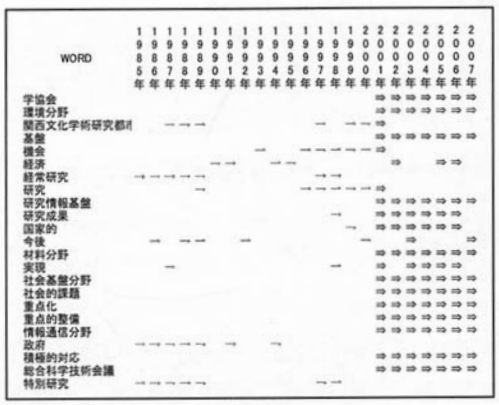

図７．単語の出現時期の分布の一部

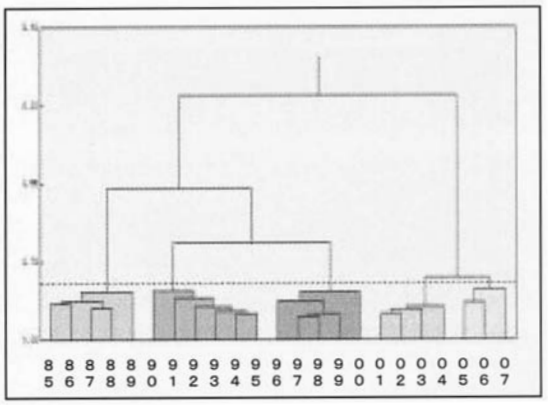

図８．ウォード法によるクラスター分析結果

ぼ一致してクラスターが形成されており、５年毎の節目があることがわかる。デンドログラムにより科学技術政策の大きな流れを可視化することができた。

６．考察並びに今後の課題

本研究では、経時的に出現した文書データの集合をもとに全体の構造を可視化する方法について検討を行った。その中でもIDFウエイトによる自己情報量は特徴を強調できることがわかった。全体の構造や流れを概観するにはクラスター分析が有効なこともわかった。しかし、自己情報量の極大値の理由を説明するのは容易ではなく、今回の研究では具体的に新語率や自己情報量の結果から原データや単語×文書マトリクスに立ちかえり、目視で確認を行った。可視化を目的とする以上、考えられる要因を自動的に提示できるシステムをめざす必要があると考える。改善方法として、原データである目次項目を構成する単語とGIDFにより数量的に評価することが考えられる。

今回は科学技術白書を対象にしたが、新聞記事のような複数パラグラフからなる文章で検証する必要がある。また、従来の方法との比較を行っていないため、今後の課題としたい。

参考文献

[1]Baeza-Yates, R. et al. Integrating Contents and Structure in Text Retrieval, *SIGMOD Record*, Vol.25, No.1, pp.67-79, March 1996

[2]Bateman, J. et al. Towards Constructive Text, Diagram, and Layout generation for Information Presentation, *ACL*, Vol.27, No.3, pp.409-449, 2001

[3]Hearst, M.A. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *ACL*, Vol.23, No.1, pp.33-64, 1997

[4]Hernandez, N. et al. What is This Text About? Combining topic and meta-descriptors for text structure presentation, *SIGDOC'03*,

[5]Kozima, H. Text segmentation based on similarity between words, *Proceedings of the 31st Annual Meeting of ACL*, pp.268-288, 1993

[6]Nakao, Y. Studies on Thematic Hierarchy Detection and Its Application for Text Summarization, A Dissertation Submitted to the Graduate School of the University of Tokyo, 2002

[7]Robertson, S. Understanding inverse document frequency: on theoretical arguments for IDF, *Journal of Documentation*, Vol.60, No.5, pp.503-520, 2004

[8]Salton, G. et al. Automatic Text decomposition Using Text Segments and Text Themes, *Proceedings of the 7th ACM conference on Hypertext*, pp.53-65, 1996

[9]Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol.28, pp.11-21, 1972

[10]相澤彰子, 語と文書の共起に基づく特徴度の数量的表現について, 情報処理学会論文誌, Vol.41, No.12, pp.3332-3343, 2000

[11]別所克人, 単語の概念ベクトルを用いたテキストセグメンテーション, 情報処理学会論文誌, Vol.42, No.11, pp.2650-2662, Nov.2001

[12]平尾努他, 語彙的結束性と単語重要度に基づくテキストセグメンテーション, 情報処理学会論文誌, Vol.41, No.SIG3(TOD6), pp.24-35, May, 2000

[13]石塚隆男, 関連性理論を用いた文章の自動要約, 情報処理学会研究報告, 2005-NL-170(18), 2005

[14]Mani, I.(奥村他訳), 自動要約, 共立出版, 2003, pp.89-104

[15]真壁肇編, 信頼性工学入門, 日本規格協会, 1985

[16]望月源他, 複数の知識の組合せを用いたテキストセグメンテーション, 情報処理学会研究報告, 1995-NL-109(18), pp.47-54, 1995