情報量に基づくテキスト・セグメンテーション

石塚　隆男

亜細亜大学経営学部

1．はじめに

本研究では、パラグラフの情報量を定義し、それに基づくテキスト・セグメンテーションの方法について検討することを目的とする。

情報洪水の時代と言われ久しいが、私たちは既存の知のごくわずかしか活用できておらず、その傾向が今後さらに加速することは疑いようがない。この問題を解決するために、情報検索システムの改善や検索されたテキストの自動要約の研究が盛んである。

情報検索の質を上げる方法として、対象テキストの文章構造に関する情報や対象分野のオントロジー（知識体系）を用いることが考えられる。また、検索結果をコンパクトに示す方法として、テキスト自動要約技術がある。さらに、検索結果全体を可視化し、利用者をナビゲートする技術も実用化されている。

利用者の情報要求が明確である場合にはこれらの方法は有用であるが、日常的に私たちは対象分野の知識や具体的な目的なしにとりあえず検索することが少なくない。検索結果の要約により確かに文章の重要点や結論は得られるが、文章の著者が何に紙幅を費やしたのか、すなわち、その文章は何について述べられているのか、を知るためには文章全体をブラウズする必要がある。書籍であれば、書名だけでは内容がわからないため、目次にざっと目を通す作業に相当する。このように、対象文章から事前知識や先入観なしに著者の言わんとするところ（＝必ずしも結論とは限らない）を読み取り、そこからさまざまな気づきや発見が得られることは活字が誕生して以来、自然とやってきた営みである。

そこで、本研究ではこうした知的活動に寄与するために文章情報を情報量に基づき、構造化することを目的としている。文章構造の手がかりとして、パラグラフ情報はきわめて重要である。

今回、パラグラフの情報量を定義し、文章の分割を試みた結果、いくつかの知見が得られたので報告する。

Text Segmentation based on Information Content
TAKAO ISHIZUKA , Faculty of B.A., Asia University

2．文章構造のモデル

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。

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

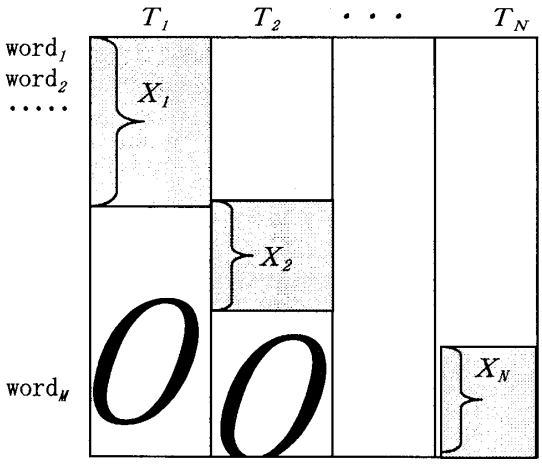

図1．単語×文書マトリクスの構造
（　　は、新語を表す）

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$M=\Sigma X_j$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。

３．パラグラフの自己情報量の定義

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。

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

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。

４．関連研究

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。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))。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。

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。

５．適用例

図２は、７パラグラフから成る日本語新聞記事の情報量の変化を示している。この記事では、第６パラグラフで話題が転換していることにより情報量が増加していることが確認された。したがって、第１・第６・第７パラグラフを抽出することにより文脈効果のある新情報だけに絞ることができると考えられる。

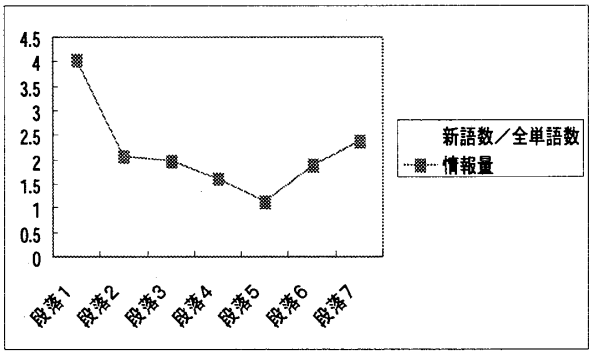

図２．記事文章のパラグラフ別情報量の変化

６．考察並びに今後の課題

テキストセグメンテーションの問題は、単語×パラグラフマトリクスの分割の問題に置き換えることができ、単語の次元とパラグラフの次元をどのように数量化するかの問題になる。たとえば、AID のような逐次分割アルゴリズムを用いたセグメンテーションも考えられ、今後の課題としたい。

参考文献

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

[2]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

[3]石塚隆男，関連性理論を用いた文章の自動要約，2005-NL-170(18)，2005

[4]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