

統計情報と文章構造特徴に基づく重要文の自動抽出

任 福継, 定永 靖史

広島市立大学情報科学部

Email: ren@its.hiroshima-cu.ac.jp

インターネットを代表とするコンピュータネットワークの発達や、CD-ROM等の大容量メディアによる出版物の普及が進み、多くの電子化された文書が流通するようになっている。本稿ではこのような文書情報へのアクセス支援のための重要文抽出方法について考察し、統計情報と文章構造特徴に基づくアプローチを提案する。重要文抽出は従来の自動要約概念と似ているが、重要文抽出では、抽出した文全体の文脈や、流暢性があまり気にしない点のみ違う。本手法では単に統計情報を用い要約文の作成を行うことではなく、文の意味的な構造情報も利用したので、従来の統計手法より、より良い重要文の抽出が期待される。この手法に基づき、重要文自動抽出実験システムを構築し、科学技術論文を用い、評価実験を行った。実験結果から、本文で提案した手法の有効性と実用性を確認することができた。

An Automatic Extraction of Important Sentences Using Statistical Information and Structural Feature

Fuji Ren, Yasushi Sadanaga

Faculty of Information Sciences, Hiroshima City University

The advent of inexpensive mass storage, particularly CD-ROM, has made possible the publication of intellectual properties such as books or journals in electronic form. Several experimental studies have been conducted to answer how to effectively access the text information. In this paper, we propose a method for extracting important sentences in articles based on statistical information and structural feature. An experimental system based on this method was constructed and an experiment on five papers was carried out. The result shows that the proposed method is effective.

1 はじめに

近年ワープロやパソコンの一般化に伴い、インターネットを代表とするコンピュータネットワークの発達や、CD-ROM等の大容量メディアによる出版物の普及が進み、多くの電子化された文書が流通するようになっている。多くの情報の中から自分にとって有用な情報を的確に選択するのは困難であるため、情報に効率的にアクセスする要求が高まっている。

情報への効率的なアクセスの手段としては、文書の検索、分類、加工の3つがあげられる[1]。

本稿では文書情報へのアクセス支援のための重要文抽出方法について考察し、統計情報と文章構造特徴に基づくアプローチを提案する。重要文抽出は従来の自動要約概念と似ているが、重要文抽出では、抽出した文全体の文脈や、流暢性があまり気にしない点のみ違う。

自動要約に関する研究手法はいくつか提案されているが、大きく分けると次の2つの種類に分類される。

一つは文章の構文や意味を考慮して要約文を作成する手法である。この手法は人が文書の要約を作成する場合の過程を模写したものともいえ、生成派とも呼ばれている。一般的に人が要約を作成する場合、文書を読んでその内容を理解し、それを再構築し直して簡潔な文章で表現するものと考えられる。すなわち「意味理解－文章再構成－文章生成」という過程を取る。このような手法では、文間の修辭的な関係に着目して「思考の流れ」として文脈を構造化する〔3〕こと、文を主張と叙述に分類して構造化を図る〔4〕こと、文同士の話題の連鎖から文章構成を捉える〔5〕こと、スクリプトなどを使用して重要箇所を抽出する〔6〕こと、中間表現形式に変換した文章に対して重要性を付与して抜粋すること、文章の表現する内容を事象という枠組みでとらえて事象を基に文の重要性を判定する〔8〕などが挙げられる。生成派では主に生成された要約の流暢さが重視されるが、処理が非常に複雑でシステム化が困難であり、十分な処理速度の確保は難しいという問題が指摘されている。また、この手法では意味解析、文生成等の技術が必要なので、現段階では思考実験段階に留まっているものも多く実用には程遠いと考えられる。

もう一つは文章の表層から得られる情報を基に、重要な文を抽出する手法である。この手法はコーパス技術を用い統計情報を有効に利用するもので、要約を文章中から重要な箇所を切り出すことと考えるため、抽出派とも呼ばれている。この場合、文に出現する各単語に重み付けを行い、その値の高い文を重要箇所とする手法が多く用いられている。重み付けには1) ヒューリスティックスを用いるもの、2) 単語頻度などを用いるもの、3) シソーラスの意味情報を用いるものなどがある。1) は、文書から得られるヒューリスティックスを用いて文の重み付けを行い重要箇所を抽出する手法である。ヒューリスティックスとしては修辭関係、タイトルに出現する語の情報、文の出現位置などがある。これらは特定の分野用に用意した知識を用いて重要箇所を抽出する手法に比べ汎用性があると言えるが、対象とする分野の変更に対してどの程度適用できるかについては調査の余地がある。2) は単語出現頻度などを用いる手法である。これには、高頻度で隣接するような語の組みに着目し、それらが多く出現する部分を抜粋するもの〔9〕や、文章の属性を調べ文脈への依存の度合いの強さで単語に対して重み付けを行う手法〔10〕等がある。3) は意味に関する統計情報を用いた手法である。これには、段落内や段落間にまたがる意味的分類の出現パターンをシソーラスを用いた分析を行い、その結果を基にチャート形式で表現するもの〔11〕や、このようなチャートを基にキーワードの自動抽出を行うもの〔12〕がある。

統計情報を用いる手法には、文章の複雑な解析を必要としないため高速に大量の文書の処理が可能である利点があるが、文の内容を把握せずに文の重要性の決定を行うために生じる問題も指摘されている。

構文や意味を用いて要約を行う場合と統計情報を用いて要約を行う場合、双方に利点、欠点があるが、現段階での実現可能性の点から考えると、統計情報を用いて要約を行うほうが優れていると考えられる。本稿で提案する手法でも文章の統計情報を用いる重要文の抽出を行う。しかし、本手法では単に統計情報を用い要約文の作成を行うことではなく、文の意味的な構造情報も利用したので、従来の統計手法より、より良い重要文の抽出が期待される。

以下では、第2節では本手法の仮定前題を述べ、第3節では意味的な構造情報の抽出方法と実験結果を記述し、第4節では重要文の抽出アルゴリズムと実験システムについて述べ、第5節では重要文抽出の評価方法及び評価結果を記述する。最後に考察と今後の課題について述べる。

2 本手法の特徴と仮定前題

統計情報と文章構造特徴に基づく重要文抽出方法とは、文中の単語の出現頻度、品詞、出現位置といった情報と、対象とする分野の文章の構造的な特徴を基に文章から重要な箇所を抜き出し、重要文を抽出するものである。

この手法の特徴としては、1) 意味解析を行わないので実用化が容易で高速な処理ができること、2) 文章の構造的な特徴情報を利用したので、抽出信頼度(精度)が向上されること、3) 対象とする分

野の文章の構造的特徴情報を追加することにより処理対象の分野の追加が簡単に実現できること、4) 要求する要約の長さを与えることによって任意の長さの要約文の作成ができることなどがあげられる。

この手法は対象とする文章にはLuhnらが提唱した前提条件 a) [13] と、我々がたてた仮説 b) が成り立つとの仮定に基づいて行っている。

- a) 1つの文献において、主題と関係の深い語や句は概して文献中に繰り返し出現する
- b) 同種類の文体において、主題と関係の深い語や句は概して同じ構造元に集中出現する

a) の仮説に基づき、まず単語の出現頻度を用いた単語の重要性判定を行う方法について検討した。この仮説に基づけば、出現頻度が高い単語を含む箇所が重要な箇所となる筈である。しかし、出現頻度だけを用いた文章の重要箇所を特定することは問題点があると考えられる。それは一般語や重要語の上位概念を含む単語は文章全体に高頻度で出現する傾向があり、単に単語の頻度のみで重要性の判定を行うと、これらの単語が重要性が高いと判定してしまうからである。また、文章中の重要語の出現分布の偏りが小さくなってしまい、出現頻度は低いものの重要性が高い単語などは、誤って重要性が低いと判断してしまう危険性が高い。一般的に、重要語となり得るものは名詞、名詞句のいずれかであると考えられる。本手法では、名詞、名詞句のみを重要語候補として重要性の判定を行うものとし、特定の範囲（本稿では章）と全体での単語の出現頻度のばらつきを考慮することによって、単語の重要性の判断を行うことにした。

b) の仮説に基づき文章構造を用いた文の重要性判定について検討した。要約を必要とする文章は一般的に長い文章であると考えられるが、長い文章には本筋と関係のない様々なトピックが含まれている。これらのトピックがノイズとなってしまうため、対象とする文章のサイズが大きくなると要約の精度が低下するという問題がある。このような問題は文、段落、節、章などの書式を含めた文章の構造を利用することにより解消できると考える。

3 科学技術論文の文章構造情報の抽出

単に統計情報を用いる文章要約においては文章の構造特徴を考慮せず、要約作成精度が高くないことが指摘されている。我々は、重要性の高い文が文章のある特定の場所に出現する可能性が高いと想定した。また、論文冒頭のアブストラクトは著者にとっての重要性の高い文章である判断から、アブストラクト中に出現する文の章ごとの出現分布と出現比率を調査した。調査に用いた論文は、情報処理系の学会誌に掲載されていたものの中から無作為抽出した20部であった。

その結果から、冒頭章もしくは最終章には重要な文があらわれる確率が高いことがわかる。この事実より、文の重要度計算の際にその構造情報を考慮する必要性があるという結論を得た。

4 アルゴリズムと実験システム

4.1 アルゴリズム

前述の調査結果を検討し、次の様なアルゴリズムにより重要文の抽出手法を提案する。

1. 文書の分類、文章へのタグ付けを行う。
2. タグ付けされた文章の形態素解析を行う。
3. 名詞と未定義語の章ごとの重要度 W_i を出現頻度をもとに計算する。
4. W_i を用い、文の重要度 SI (Sentence Importance)を計算する。
5. SI と文章構造情報より得た各文の構造的重み SW (Sentence Weight)をもとに、各文の主題重要度 TI (Theme Importance)を求める。
6. 要求された要約長さより、 TI を用いて要約文 SS (Summary Sentence)を抽出する。

4.2 システム構成

本システムはテキスト前処理部、形態素解析部、統計情報抽出部、単語重要度計算部、文重要度計算部、構造的重みにより文重要度の調整よりなる。

a テキスト前処理

入力された文章の分類と、文章構造へのタグ付けを行う。今回は論文のみを選択し、文構造へのタグ付けは手作業で行った。文構造へのタグ付けは、章の分け目に任意の文字列で括った章の番号の数値を打ち込むことで行った。

b 形態素解析

本研究では形態素解析にJUMANを使用した。

先に説明したように重要語となり得るものは、名詞、名詞句のいずれかである。JUMANは全・半角アルファベット、カタカナ、辞書中に含まれない語等に対して、未定義語との判定を行う。しかし、論文中で新たに定義された単語や、新規に作られた単語などは辞書に含まれていない場合未定義語となることが多い。また、アルファベットやカタカナの単語等も未定義語となるが、これら未定義語の中には文中で重要意味を持つ場合が多く、重要語である可能性は高い。このため、本手法では文章の形態素解析後に名詞、未定義語に対してのみ単語の重要度計算を行うこととした。

c 単語情報抽出

形態素解析された結果をもとに形態素、文、章、文書の情報の抽出を行う。入力文書の章の数、文の数、形態素数、名詞と未定義語数等の情報を得る。

d 単語重要度計算

章ごとの名詞、未定義語の重要度計算を行う。以下に単語重要度計算のための計算式を定義する。

$$W_i = \alpha_1 \frac{\ln_i}{m} + \alpha_2 \frac{\ln_i}{n} + \alpha_3 \frac{\ln_i}{\ln_i}$$

W_i : 単語iの重要度

$\ln_i$: 文章全体での単語iの出現数

$\ln_i$: 章での単語iの出現数

m : 文章全体での名詞と未定義語の出現数

n : 章での名詞と未定義語の出現数

α_i : 調整用パラメータ ($i=1,2,3$) (ただし $\alpha_1 + \alpha_2 + \alpha_3 = 1.0$ かつ $\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0$)

$\ln_i/m$ は、文章全体での単語の出現比率を表す。これにより、全体で高頻度に出現する単語の重要度が求められる。 $\ln_i/n$ は章での単語の出現比率が求められる。これにより、章で高頻度に出現する単語の重要度が求められる。 $\ln_i/\ln_i$ は単語のその章での出現比率が求められる。文章全体で平均的に出現している語は、ここの値が小さくなり、特定の章に片寄って出現している単語はここの値が大きくなる。これにより、高頻度で文章全体で出現すると思われる一般語に、高い重要度を与えることを避けられると考えられる。

e 文重要度計算

得られた単語重要度 W_i を用い、文の重要度 SI を求める。文重要度を得るために以下の様な計算式を定義する。

$$SI_j = \frac{\sum_{i=1}^k W_i}{\sqrt{k}}$$

SI_j : 文 j の重要度

k : 文中の名詞、未定義語の総数

これは文に含まれる単語の重要度を合計し、文中の単語の総数の平方根で除したものである。文中の単語の数ではなく単語の平方根で除している理由は、文中に含まれる単語数により重要度が受ける影響を小さくするためである。この式により、文中に含まれる単語の重要度をもとに文の重要度を決定できる。

f 構造的重みにより文重要度の調整

得られた文重要度 SI_j を用い、文の構造的重みによる修正を行った文重要度 TI を求める。本稿では文章を、冒頭章、中間章、最終章の3つに分類した。文重要度を得るために以下の様な計算式を定義した。

$$TI_j = f(n) \times SI_j$$

$$n = \{1, 2, 3\} \quad f(1) = \beta_1 \text{ (冒頭章)} \quad f(2) = \beta_2 \text{ (中間章)} \quad f(3) = \beta_3 \text{ (最終章)}$$

$$\text{ただし } \beta_1 + \beta_2 + \beta_3 = 1.0, \beta_1 > 0, \beta_2 > 0, \beta_3 > 0$$

先に得られた文の重要度 SI_j に、出現している章による構造的な重みを考慮して、文の重要度の修正を行うことによって、最終的な文の重要度 TI_j を求める。

g 重要文の抽出

文の重要度 TI_j を基に重要文の抽出を行う。まず得られた TI_j により文全体を文の重要度の高い順にソートし、設定した文の数に見合った数の文を重要文として出力する。

5 実験及び評価

本手法の有効性を確認するために、前章で述べた要約システムのプロトタイプを作成し、重要文の抽出実験を行った。システムは言語C++で作成し、FreeBSD上に実装した。

5.1 実験

実験では日本語で記述された科学技術論文5本を用いて行った。

下記の4つの組みの単語の重要度計算式のパラメータ $\alpha_1, \alpha_2, \alpha_3$ を用い、4回ずつの抽出実験を行った。

セット1 : $\alpha_1 = 0.33, \alpha_2 = 0.33, \alpha_3 = 0.33$

セット2 : $\alpha_1 = 0.20, \alpha_2 = 0.20, \alpha_3 = 0.60$

セット3 : $\alpha_1 = 0.20, \alpha_2 = 0.60, \alpha_3 = 0.20$

セット4 : $\alpha_1 = 0.60, \alpha_2 = 0.20, \alpha_3 = 0.20$

なお、構造的重みを表示するパラメータ $\beta_1, \beta_2, \beta_3$ は、構造的特徴の調査の結果を基に

$$\beta_1 = 0.55, \beta_2 = 0.10, \beta_3 = 0.35$$

を採用した。

本稿では、重要文の抽出文数は入力した論文の長さに関係無く一定とし、システムが重要度が高いと判定した順に10文および20文とした。

5.2 実験の評価

今回は、評価実験に用いた論文の著者自身に要約として作成された文が重要であるか否かの判定を行ってもらった。著者自身に判定を行ってもらうことで、第三者に判定を行ってもらった際に起こりうる重要性の判定の誤りを避けることができ、信頼度の高い判定を行なえると考えたためである。

評価は以下の5段階で行った。

重要	やや重要	普通	重要ではない	不要
5	4	3	2	1

抽出信頼度の定義

評価指標として抽出信頼度ARを用いるが、下記のように定義されている。なお、評価結果を表1に示した。

$$AR(\%) = \frac{\sum_{i=1}^n P_i}{n} \times \frac{100}{5}$$

Pi = 要約文の評価得点

n = 要約文数

表1. 抽出信頼度

	セット1		セット2		セット3		セット4	
	20文	10文	20文	10文	20文	10文	20文	10文
著者1	62	56	56	52	64	64	66	60
著者2	59	78	56	72	59	70	57	78
著者3	61	58	63	62	61	58	61	58
著者4	74	64	72	62	71	74	71	74
著者5	61	70	56	62	60	70	59	68

要約処理時間

使用機材：GATEWAY2000P5-200,Pentium200,RAM64MB

使用OS：FreeBSD2.2

上記の環境における要約処理時間は、表2のようになっており、作成したシステムは十分実用的な速度で動作していることが分かる。

表2. 処理時間

	論文容量(byt)	形態素解析済み容量(byt)	要約処理時間
著者1	19142	119565	3秒
著者2	32137	208145	6秒
著者3	19319	181495	3秒
著者4	8876	100880	5秒
著者5	26551	171180	3秒

5. 3 考察

著者が不要と判定した文を調べた結果、例示を行う文や語の説明を行う文が複数みられた。これらの文には、「示す、説明する、例えば」等といった特徴的な表現が含まれる傾向があった。このような表現を持つ文を「不要文」と定義し、不要文を削除することによって、抽出精度が改善されるかを判定するため、ハンドシミュレートを行った。その結果を表3、4に示す。

本稿では、ある単語が特定の範囲での出現頻度の方が他の範囲での出現頻度に比べて著しく高い場合、その単語はその範囲内で重要性が高い、との仮説を立てて単語の重要度計算に用いた。単語の出現章によるばらつきに高く重み付けを行った結果と、それ以外の場合との結果を比較してみたが有為な差

は得られなかった。また、実験結果より、ばらつきを重視し過ぎた場合、1度しか出現していない単語の重要度が極端に高くなるという問題があることも明らかになった。

表3. 不要文を削除した抽出信頼度の変化 (重要度上位10文)

シミュレート	セット1		セット2		セット3		セット4	
	前	後	前	後	前	後	前	後
著者1	56	56	52	52	64	64	60	60
著者2	78	78	72	72	70	70	78	78
著者3	58	64	62	68	58	64	58	64
著者4	68	74	64	64	74	80	74	74
著者5	70	70	62	62	70	70	68	68

表4. 不要文を削除した抽出信頼度の変化 (重要度上位20文)

シミュレート	セット1		セット2		セット3		セット4	
	前	後	前	後	前	後	前	後
著者1	62	62	56	56	64	64	66	66
著者2	59	59	56	56	59	59	57	57
著者3	61	67	63	69	61	70	61	67
著者4	74	76	72	74	71	71	71	74
著者5	61	61	56	56	60	60	59	59

単語の重要度計算の段階で、不適切な切り分けをされた単語が見られた。本手法では単語の品詞情報などは形態素解析を行うことで得ており、形態素解析の誤りによる重要語選定の失敗が避けられない。また、JUMANが形態素解析を誤った語のほとんどが名詞と判断されており、名詞、未定義語についてのみ重要度計算を行う本手法では少なからず影響があると思われる。今後、形態素解析誤りの与える影響についての調査を行う必要がある。

JUMANの判定する名詞には普通名詞、サ変名詞、固有名詞、地名、人名、数詞、形式名詞、副詞的名詞、時相名詞がある。本稿ではキーワード抽出を行わず名詞、未定義語すべてに対して重要語候補とし、重要度の計算を行った。しかし、名詞のすべてが重要語とは限らず、重要語となりやすい名詞と重要語になりにくい名詞があると思われる。今後、名詞の活用型の違いを考慮に入れ再度評価実験を行いたい。

6 おわりに

本稿では、統計情報と文章構造特徴に基づく重要文抽出手法を提案し、実験システムを構築し、評価実験を行った。実験結果より、本手法の有効性を確認することができた。しかし、実用化システムとしては、まだ改善すべき問題が多数あることも明らかになった。

文書の内容をあらわす重要語は語長が長いことが多い。これは、語長が長いほど語の意味が具体化されるためである。しかし本手法では、単語の切り分けを形態素解析で行っているため、複合語は複数の単語に分割されてしまっており、内容把握に必要な複合語の獲得には不向きである。適切な隣接語を併合することでこの問題は解決できるが、今度は一般的に語長の長い語は出現頻度が低いことが多いため、重要語の出現頻度が少ない場合に高い重要度を付与できないという問題が生じる。

また、出力段階で文章内の比較的重要度が低いと考えられる語や不要と思われる文の削除は実現すべき課題ということが確認できた。

謝辞

本研究の一部は文部省科研費により行われた。本研究にあたり、熱心に議論して頂いた、広島市立大学情報科学部自然言語処理学講座の各位に感謝いたします。また、形態素解析システムJUMANを提供していただいた京都大学の長尾真先生、奈良先端科学技術大学院大学の松本佑治先生および開発グループの方々、評価実験用に論文を提供し要約文の評価を行っていただいた方々に感謝いたします。

参考文献

- [1] 自然言語処理、長尾真編、岩波書店、1996.
- [2] 要約文の表現類型－日本語教育と国語教育のために－、佐久間まゆみ編、ひつじ書房、1994.
- [3] 文脈構造の解析、小野頭司 浮田輝彦 天野真家、情報処理学会研究報告NL, 70巻, 1989.
- [4] 著者の主張に基づく日本語文章の構造化、福本順一、情報処理学会研究報告NL, 78巻, 1990.
- [5] 文章の表現形式に基づいた要約文章の作成に付いて、田村俊哉 田村直良、情報処理学会研究報告NL, 92巻, 1992.
- [6] 要約過程の形式化と実現について、田村直良、人工知能学会誌, 4巻, 1991.
- [7] 要約支援システムCOGITO、安原宏 小松英二 日比孝 下等安彦、情報処理, 第30巻10号, 1989.
- [8] 事象解析による要約情報の抽出、稲垣博人、言語理解とコミュニケーション, 91巻, 1991.
- [9] キーワード密度方式自動抄録の改良、鈴木康弘 枅内香次、情報処理学会論文誌, 第29巻3号, 1988.
- [10] 文脈依存の度合いを考慮した重要パラグラフの抽出、福本文代 福本純一 鈴木良弥、自然言語処理, 第4巻2号, 1997.
- [11] 語彙の結束性に着目した文章抄録法の提案、佐々木一朗 増山繁 内藤昭三、情報処理学会研究報告NL, 98巻, 1993.
- [12] 語の意味分類を考慮したキーワード抽出の試み、鈴木斎 増山繁 内藤昭三、情報処理学会研究報告NL, 98巻, 1993.
- [13] "The Automatic Creation of Literature Abstracts.", Luhn, H.P IBM journal, 2(1), 1958.
- [14] 人間の重要文判定に基づいた自動要約の試み、野元忠司 松本祐治、情報処理学会研究報告NL, 120巻, 1997.
- [15] 日本語形態素解析システムJUMAN version 3.4 マニュアル、黒橋禎夫 長尾真、京都大学工学部 奈良先端科学技術大学院大学, 1997.
- [16] 文章内構造を複合的に利用した論説分要約システムGREEN、山本和英 増山繁 内藤昭三、自然言語処理, 第2巻1号, 1995.
- [17] 特徴的表現を利用した特許抄録作成法の検討、原正巳 木谷強 江里口善生、情報処理学会研究報告NL, 100巻, 1994.