自然言語処理 135－14
（2000．1．28）

文章要約のための特徴キーワードの発見による重要文抽出法－展望台システム－

A System for Making Summary by Extracting Key Sentences Using feature keywords -A Panoramic View System-

砂山 渡 Wataru Sunayama
大阪大学大学院基礎工学研究科システム人間系専攻
Dept. of Systems and Human Science, Graduate School of Engineering Science, Osaka University
sunayama@sys.es.osaka-u.ac.jp

谷内田 正彦 Masahiko Yachida
大阪大学大学院基礎工学研究科システム人間系専攻
Dept. of Systems and Human Science, Graduate School of Engineering Science, Osaka University
yachida@sys.es.osaka-u.ac.jp

Keywords: Summary, Keywords Extraction, Extraction of Key Sentences, Co-occurrence of keywords,

Summary

There are so many resources of electronical documents. However, we cannot access the all documents and we cannot examine each documents precisely. Therefore summaries are available for the selections of the documents. A system for making summary is Panoramic View System(PVS) which extracts key sentences from a document by the original method. PVS extracts keywords from the top of the term-frequency mountain. Many ordinary systems extract keywords by high frequency. However, PVS also extracts low frequency keywords based on the co-occurence of the keywords in sentences. Such keywords feature each sentence and are useful for extractions of key sentences.

1. はじめに

近年，インターネットの利用が盛んになるにつれ，扱われる情報量が大幅に増えてきている．情報の伝達に用いられる文書も電子化が進み，E-mailをはじめ，Webページ，ネットニュース，電子図書館などさまざまな手段を通じて文章に接し，情報を獲得する機会が増えている．しかしその情報量はあまりに膨大であり，時間や心身の制約によって，全ての情報に目を通すことは非常に困難である．

最近，そのような多くの情報の中から必要な情報を素早く獲得するための研究が，情報検索などの分野で盛んに行なわれている．最も代表的な情報検索の手段としてはサーチエンジンが挙げられ，ユーザが入力した検索語を含むページを出力する際に，各検索語の各Webページ中での重要度を独自に決めて順位付けの後に出力される*1．すなわち，各Webページのキーワードとユーザが入力した検索語とが数多く一致していれば上位に出力される評価関数となる．それゆえ文章の中で重要な部分（キーワードや重要文）はどこかを見つけ出すことには，検索のランキングを改善する上で意味がある．

また，現在サーチエンジンの出力によく見かける短い概要は全て，検索された各Webページの冒頭から取り出されている．ニュース記事のように冒頭文に意味のある内容が含まれているという制約は一般には満たされないため，結局はそのページまでハイパーリンクを辿った上で，入力した検索語がどのように使われているかを実際に確認することもしばしばである．この原因は，文章中で重要な部分を特定できないことによっている．あらかじめ重要文を抽出しておくことで，検索語を含む重要文を優先して概要に用いることによっても，検索の能率が上がると考えられる．

検索の能率を上げる従来研究として，Webページに含まれるキーワードをユーザに提供して検索を支援するシステム[砂山 99]がある．これは各Webページから頻度によって順位づけされたキーワードを用いて，検索語にマッチした多くのWebページに共通に現れるキーワードをユーザに提供している．この

*1 たとえば，「タイトルに含まれる」や「フォントが大きい」などである．

あらかじめ各ページに与えておくキーワードは，各Webページ中で重要とされる文に含まれていなければ，検索支援としての効果は十分ではないだろう．

従来の重要箇所（段落，文，節など）の抽出およびその連結による要約生成においては，文章に含まれる各文を評価した上で，評価値の高い文を抽出するという方法が採られている．各文の評価には，文に含まれる単語のキーワードとしての評価値を単語の出現頻度や，テキスト固有に出現する単語の評価を組合せたTFIDF法 [Salton 97] による値などで与え，その評価値の和で文の重要度を表す方法が広く用いられている．また，文間の関連度に基づいて多くの文と関わりがある文を選ぶ手法もある [奥村 99]．文と文は，両者に出現する同じ単語もしくは同一概念を表す単語の出現によって結びつけられ，他の文と多くの結び付きをもつ文が重要であるとされる．この考え方は語彙的連鎖 [Morris 91] に代表され，情報検索やキーワード抽出にも応用されている．そのほか，文章中における文の位置情報 [Brandow 95] や接続詞など特定の単語を手がかりとする方法 [任 98] などがある．

本稿で提案するシステムにおいては，キーワード抽出に基づく重要文抽出を行なう．従来からの頻度によるキーワードを手がかりとして，文章の主題となるキーワードを単語間の関連に基づいて抽出し，低頻度のキーワードの中からも文章の主題に合致しかつ文章を特徴付けるキーワードを取りだす．

以下の**2.**で提案するシステムの概要とアルゴリズムの詳細について述べ，**3.**で文章から重要文を抽出する実験を行なう．**4.**で実験結果に対する考察および評価を行ない，**5.**で結論と今後の展望について述べる．

2. 展望台システム

2·1 展望台システムの概論

真ん中に大きな山がある町を想像して欲しい．町の特徴をとらえて宣伝するためには如何なる方法が必要だろうか．山以外の町中の一点から町全体を眺めようとしても，視界には山ばかりが強調され，それ以外に周辺の建物がいくつか見えるだけであろう．特に山の向こう側なんて何があるか検討もつかない．そこで，我々のシステムは山に登る．山頂にある展望台ならば町全体を見渡すことが可能である．町全体を見渡す中で，この町を特徴付ける劇場や博物館，公園など，山の麓からでは気づきにくい場所までも，

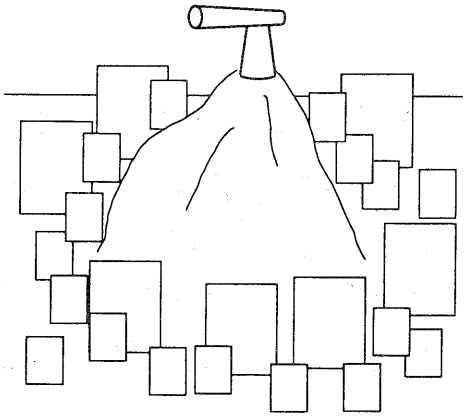

図1 展望台システム概観

しっかりと捉えることができるであろう．

本システムでは一つの文章を，単語の出現頻度によって表される山を含む町とみなす．頻度の高い単語による山があり，頻度の低い単語は民家や学校，病院となる．文章を読む際には，繰り返し現れる単語は読者の印象に残り文章の中の大きな特徴として捉えられる．しかし，実際には見落とされやすいけれども文章の特徴付けをしている頻度が低いけれども重要な単語が存在するのである．本稿では，そのような文章の特徴づけを行なっている単語を探して評価することによって，文章から重要文を抽出する方法について述べる．

2·2 本システムの目標：要約とは？

本システムが出力する重要文とは，元の文章の中で重要と評価された文そのものであり，これは「抄録*2」に当たる．この抄録をまとめる方法は，目的や用途によって異なると考えられるが，文章の全範囲のうちどの程度をカバーするかによって次の3種類に分けられる．それらはすなわち，

1. 縮約：文章全体を縮尺してまとめる [文章の全体]
2. 概要：(要点をかいつまんだ) だいたいの内容 [文章の一部]
3. 要約：文章の要点を捉えて短く纏めたもの [文章のポイント]

の3つである．単純に文章全体から重要文を抽出した場合，その抽出する重要文の全文に対する割合によって，1から3が区別できるとも考えられる*3．しかし，単純に抽出する文の割合を変化させても，文

*2 （文章の大事な部分を）抜き出したもの
*3 抄録の読みやすさ [難波 99] に関する議論の余地はある．

章が長くなるにつれて，取り出される重要文の文章内での位置に偏りが生じてしまう．そのため，文章の縮約を考える際には，文章の分割 [平尾 99] などの手法を組み合わせ，段落ごとの長さに応じて抽出する重要文の割合を決定するなどの工夫が必要となる．特に文章の分割が元からなされている，もしくはなされたと仮定した場合の文章要約作成の手順は図 2 のようになる．

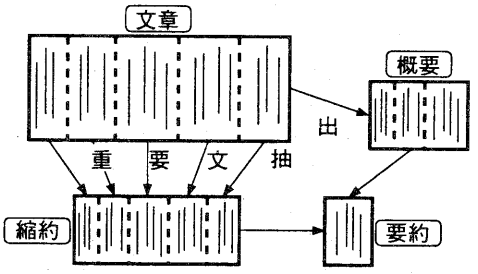

図 2　要約生成の流れ

図 2の各図内の点線は段落の切れ目を表している．文章からは，重要文抽出によって概要を生成し，段落ごとの重要文をもとに縮約が生成することができる．その後に，縮約と概要を組み合わせることによって，文章全体の流れと，文章中の重要な部分との組合せによって，適切な要約が生成できると考えている．

2·3　本システムの扱う文章

本節では，本システムが扱う文章の性質について説明する．入力となる文章の性質には，テキストの長さ，ジャンル，分野，単一／複数テキストの別などがある [奥村 99]．テキストの長さについては，1 文や 2 文などの極端に短い場合を除いて特に問わない．頻度などの統計情報を用いる場合，抽出するキーワードや文の精度を得るために，ある程度の長さを文章に求めるのが通常であるが，本システムの狙いとして，単語間の関連に基づいて短い文章からも正確に重要な単語や文を抽出することを目指している．

また文章のジャンルや分野に関しても，ジャンルや分野に応じた知識や，意味処理を行なわないために特に指定はない*4．

最後に，扱うテキストは単一のテキストとする．しかし文章全体からの重要文の抽出を考える際には，本システムは，文の出現順序に関する情報を用いないために，与えられた全てのテキストを並べてできる 1 つのテキストに対して，本システムが用いられる可能性はある．

*4　形態素解析を用いているため，日本語で辞書に含まれている単語が使われてあることと，文の切れ目である読点が文章に含まれていることが望ましい．

2·4　抄録生成アルゴリズム

本節では，入力となる文章から重要文を抽出して抄録を生成するアルゴリズムについて説明する．各ステップの括弧内に 2·1に基づく物理的意味を示す．

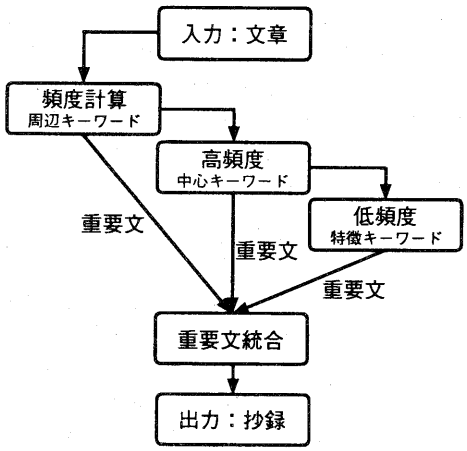

図 3　抄録生成アルゴリズム

ステップ 0：キーワード候補の準備
(町作りのための整地)

文章のキーワードの候補として，予め形態素解析 [Chasen] によって，名詞（普通名詞，サ変名詞，固有名詞，地名，人名），動詞，形容詞*5を取り出しておく．

ステップ 1：周辺キーワードの抽出
(頻度による単語の山と町を生成)

まず，単純に単語の出現頻度によって，キーワードを抽出する．このキーワードを周辺キーワードと呼ぶ．

$$key1(w) = frequency(w) \qquad (1)$$

周辺キーワードの数は，文章に含まれる名詞の種類数の２０％を目安として，最大で２０個を取る*6．高頻度によって単語に高評価を与える多くの手法があるように，複数回用いられる単語の重要性は再び言うまでもない [Luhn 58]．しかし，頻度のみでは各単語が文章の中でどのように現れ，どのような役

*5　「する，なる，ある，よる，いう，ない」を除く．

*6　２０％という値は，予備実験から得られた文章中で複数回出現する単語の割合の平均値である．同頻度の単語はすべて選択するという条件の下で２０％に最も近くなるしきい値を設定した場合，全ての周辺キーワードが複数回出現する単語となった．

割を果たしているかを知ることができないので，単語間の関連をもとに重要なキーワードを探す操作が必要となる．

ステップ 2：中心キーワードの抽出 (展望台の建築)

次に，周辺キーワードが現れた時に現れやすい単語として中心キーワードとして抽出する．これは，文章の頻度による多くのキーワードを土台として，文章の主張を表すキーワードが存在する [大澤 99] という考えに基づいている．すなわち，文章中の各単語に次の評価関数で与えられる評価値を与え，その評価値による上位（単語種類数の 4 %で最大 5 個）を中心キーワードとする．周辺キーワード集合 G，単語 w を含む文の数を $n(w)$ とするときの単語 w の評価値 $key(w)$ は次式となる．

$$key(w) = \prod_{g \subset G} \frac{n(w \cap g)}{n(g)} \qquad (2)$$

この式は，周辺キーワード g が出現した時に単語 w も同時に出現する条件付確率を，すべての周辺キーワードに関して積算している．そのため，頻度による周辺キーワードと同時に出現する割合が高いキーワードは，特に頻度が高いキーワードに限られる．しかし，単語の文内における共起関係をもとにしているために，必ずしもすべて頻度順に選ばれるわけではない．

ステップ 3：繰り返し中心キーワードの抽出 (展望台の補強)

前ステップでは全ての周辺キーワードを手がかりに中心キーワードを探していたが，周辺キーワードよりも文章の主題を直接表す中心キーワードを手がかりとすることで，取り損ねている文章の核となるキーワードを補う．そこで，中心キーワード集合 S が現れたときに，同時に現れやすいキーワードを求める．すなわち，全ての単語に再び次の式による評価値を与え，その評価値による上位（単語種類数の 4 %で最大 5 つ）に含まれかつ，まだ S に属さないキーワードを新たな中心キーワードとして S に加える．

$$key2(w) = \prod_{s \subset S} \frac{n(w \cap s)}{n(s)} \qquad (3)$$

この評価関数は，先の評価関数の周辺キーワード集合を中心キーワード集合と置き換えたものに一致する．このステップ 3 を新たな中心キーワードが追加されなくなるまで繰り返す．

ステップ 4：特徴キーワードの発見 (展望台からの眺めによる発見)

最後に文章の主題となっている中心キーワードの流れに沿い，かつ文章を特徴付けているキーワードを抽出する．そのための単語の評価値を次式で与える．

$$key3(w) = \prod_{s \subset S} \frac{n(w \cap s)}{n(w)} \qquad (4)$$

すなわち，中心キーワードが出現する時にのみ出現する単語を探す目的で，どれだけ多くの中心キーワードと密接に関係しているかを表す評価値を各単語 w に与える．文章の中で重要な役割を果たすキーワードは，高頻度のキーワードだけではない．文中に数少なく現れるキーワードの内にも，本文を表す上で重要なキーワードが存在している．このような特徴キーワードは特に頻度が低くなる傾向にあり，中心キーワードが多く現れる文でしか使われていない単語が，特徴キーワードとなる．すなわち，文章の主題となる中心キーワードから逸れない範囲で，他の文には現れていない単語を高く評価することに相当し，単純に文単位で tfidf を行なった時に比べて，文章の主題に関して一貫性のある文からのキーワードが得られる．

ステップ 5：文章中の各文を評価（地域の評価）

文章中の各文に文中に含まれる単語の評価値の総和を各文の重要度として与える [Zechner 96]．各文 T には式 (5)から式 (7)の 3 つの評価値を与える．ただし単語 w の評価値 $key1(w)$ が，全単語の式 (1)による評価値の平均を下回る場合には，$key1(w) = 0$ とする．また，$key2(w), key3(w)$ についても同様の操作を行なう[*7]．

$$sentence1(T) = \sum_{w \subset T} key1(w) \qquad (5)$$

$$sentence2(T) = \sum_{w \subset T} key2(w) \qquad (6)$$

$$sentence3(T) = \sum_{w \subset T} key3(w) \qquad (7)$$

ステップ 6：重要文の決定（宣伝地域の特定）

式 (1)から式 (4)による各文の評価値を元に，重要文を総合的に決定する．それには，まず各文に式 (5)から式 (7)のそれぞれの値に基づいて，評価値の高い順に順位をつける．すなわち，各文 T は式 (5)による順位 $rank1(T)$，式 (6)による順位 $rank2(T)$ と式 (7)による順位 rank3(T) の 3 種類の順位をもつ[*8]．その順位づけの後に，式 (8)の各文の順位の合計をその文の評価値として与える．

*7 この操作は，長い文ほど自然に評価値が高くなることを抑えると共に，有効な単語とそうでない単語とを明確に区別するために行なっている．また，有効な単語を逃さないで十分に含むように，各評価関数による評価値の平均をしきい値として設定した．

*8 同じ評価値をもつ場合は同順位となる．

表1 周辺キーワード（上位）

単語	評価値（頻度）
キーワード	105
文	76
文章	73
重要だ	45
抽出	35
単語	35
中心	28
システム	26
含む	23
頻度	20

表2 中心キーワード（上位）

単語	評価値	頻度
＊文	3.05	76
＊キーワード	2.23	105
＊文章	2.22	73
＊重要だ	2.07	45
＊抽出	1.33	35
システム	0.97	26
頻度	0.87	20
含む	0.83	23
中心	0.74	28
＊単語	0.67	35

$$value(T) = rank1(T) + rank2(T) + rank3(T) \quad (8)$$

この，式(8)の値に基づいて，用途に応じて必要な数[*9]の重要文を出力とする．

3. 実　　験

本手法に基づいて，文章に含まれる各単語（名詞）に評価値を与えて，重要文を抽出する実験を行なった．実験環境は，Pentium II 350MHz(256MB) OS-Linux であり，プログラムはC言語で書かれている．実験に用いたテキストは，本稿（本章を除く 153 文 492 種 1849 個の単語）である．

まず，表1の頻度に基づいて上位 20 個の単語を周辺キーワードとして取り出す（2·4のステップ1）．

次に，周辺キーワードが現れた時に出現しやすいキーワードを，中心キーワード（5つ）として取り出す（ステップ2）．また，取り出された中心キーワードを元に，再帰的に中心キーワードを集める操作を行ない（ステップ3）中心キーワード集合（表2の＊印の単語）が生成された．

得られた6個の中心キーワードを元に，中心キーワードが出現する時にのみ用いられやすいキーワードを特徴キーワード（表3）としてその評価値を得る（ステップ4）．

次に，各単語の評価値に基づいて各文に評価値を与える（ステップ5）[*10]．最後に，各キーワード抽出法による文の評価値による順位を元に重要文を出力する（ステップ6）．本稿の上位5文による抄録は以下のようになった（出現順に表示）．太字の単語が主な特徴キーワードであり，括弧内に式(8)の値（合計＝周辺＋中心＋特徴）がある．

表3 特徴キーワード（上位）

単語	評価値	頻度
統計	34.01	1
狙う	34.01	1
精度	34.01	1
正確だ	34.01	1
目指す	17.74	2
特徴づけ	15.12	1
動作	15.12	1
質	15.12	1
技術	15.12	1
悪い	15.12	1
本稿	12.81	2
向上	10.92	2

1.(24=3+7+14)「従来からの頻度によるキーワードを手がかりとして，文章の主題となるキーワードを単語間の関連に基づいて抽出し，低頻度のキーワードの中からも文章の主題に合致しかつ文章を特徴付けるキーワードを取りだす.」

2.(11=8+2+1)「頻度などの統計情報を用いる場合，抽出するキーワードや文の精度を得るために，ある程度の長さを文章に求めるのが通常であるが，本システムの狙いとして，単語間の関連に基づいて短い文章からも正確に重要な単語や文を抽出することを目指している.」

3.(15=5+3+7)「すなわち，文章の主題となる中心キーワードから逸れない範囲で，他の文には現れていない単語を高く評価することに相当し，単純に文単位で tfidf を行なった時に比べて，文章の主題に関して一貫性のある文からのキーワードが得られる.」

4.(10=4+1+5)「一つのキーワード抽出法によって重要文抽出を行なうシステム，あるいはさまざまなキーワード抽出法を単純に組み合わせてその多数決によって重要文を抽出するシステムの場合，取り出される重要文には一貫性がないことが多い.」

5.(20=7+4+9)「現在はこのキーワード集合を観点とした重要文を抽出しているが，中心キーワードとして別の観点に基づくキーワード集合を用意すれば，異なる観点に基づく重要文抽出にも

*9 最重要文が必要な場合は1つ，10％の抄録に用いる場合には全文数の10％に当たる数の文を取り出す.

*10 評価値のしきい値となる平均は周辺，中心，特徴キーワードの順にそれぞれ 3.4, 0.32, 2.36 となった

応用が可能である.」

3·1 実験結果の解釈

中心キーワード集合として得られた6個の単語「文, キーワード, 文章, 重要だ, 抽出, 単語」は, この実験例においては頻度順であるが, 本稿のタイトルが「特徴キーワードの発見による重要文抽出」であることから, 文章の主題として適切である. ここでもし「特徴」という言葉も主題となる中心キーワードに含めたいと筆者が考える場合には, 4·2で述べる「観点」として与えることや, 5.で述べるように文章自体を「特徴」という言葉がキーワードとして得られるように改善を図ることが考えられる.

また, 中心キーワードの評価値は頻度順とはなっておらず, 単語間の関連を元に評価値が与えられている. この中心キーワードの評価値による重要文は上記出力の順位は, 1,2,3,4,7 となっており, 本実験結果においてはその効果が十分に現れている. また, 抄録の各キーワードに基づく順位にはばらつきがあり, 周辺キーワードと特徴キーワードを中心キーワードによって結び付け統合している様子がうかがえる.

最後に, 肝心の抄録の内容についてであるが, 主題を表す中心キーワードを多く含んでいるとともに, 周辺キーワードによる頻度が高いキーワードを含みつつ, 中心キーワードとのみ出現する低頻度の特徴キーワードを含んだ文が選ばれている.

4. システム評価

4·1 キーワード抽出法の間の相関関係

次の4つのキーワード抽出法によって抽出される重要文の間の相関関係を調べる実験を行なった. 実験に用いた文書は Web 上の芸能関係のニュース記事5526件[*11]である.

1)周辺キーワード
2)tfidf によるキーワード
3)中心キーワード
4)特徴キーワード

まず, 1) と 2) の比較を行なった. その結果, 最重要文の 82 %が一致し, またそれぞれの出力の上位5文を比較したところ, 平均で 4.3 文が一致しており, 9割の文書で 4 文以上が一致していた. この結果の理由としては, 扱った文書が短いこととは別に, ニュース記事は今までは無かった事や変わった事新しい事を記事にするために他の文書にはないキーワードがそのままその文書において多く使われているためだと考えられる.

次に, 本システムで用いる 2)3)4) の3種類のキーワードによる重要文抽出の結果を比較した.

2つのキーワード抽出法 A, Bによる重要文の一致度 $Similar(A,B)$ を式 (9)で定める. ただし, n は抽出する重要文の数であり, $Same(A,B)$ は A と B で一致する文の数の平均である. また, 3つのキーワード抽出法による一致度も同様に式 (10)で定める.

$$Similar(A,B) = \frac{Same(A,B)}{n} \quad (9)$$

$$Similar(A,B,C) = \frac{Same(A,B,C)}{n} \quad (10)$$

この一致度をまとめた表が表4と表5である. 表4はキーワードとして名詞のみを扱った場合の結果であり, 表5はキーワードとして名詞と用言の両方を扱った場合の結果である.

表4 抽出する重要文と一致度の関係 (名詞)

要約文の数	周中	周特	中特	周 tfidf	周中特
1	0.699	0.599	0.842	0.822	0.587
2	0.695	0.606	0.820	0.830	0.580
3	0.717	0.647	0.821	0.839	0.610
4	0.742	0.684	0.826	0.853	0.642
5	0.768	0.718	0.836	0.866	0.674
6	0.794	0.749	0.851	0.875	0.706
7	0.817	0.779	0.866	0.887	0.738

表5 抽出する重要文と一致度の関係 (名詞・用言)

要約文の数	周中	周特	中特	周 tfidf	周中特
1	0.629	0.577	0.800	0.753	0.525
2	0.650	0.560	0.752	0.754	0.517
3	0.690	0.567	0.713	0.766	0.525
4	0.717	0.569	0.685	0.783	0.523
5	0.742	0.577	0.672	0.802	0.530
6	0.766	0.589	0.671	0.817	0.541
7	0.784	0.609	0.683	0.834	0.559

この結果, 中心キーワードと特徴キーワードの一致度が高い値を示している. これは文章が短いために中心キーワードを含む文の絶対数が少なく, 中心キーワードに関連して得られる特徴キーワードを含む文に一致したためと考えられる. また名詞のみの場合に比べて, 用言をも考慮した場合に全体に一致度が下がっている. これは, 一文から抽出されるキーワードの数が増えたことによって, 各キーワードへの評価が分散したためと考えられる.

そこで, より長い文章に対して同様の実験を行なった. 文章には, 27 の書簡 (平均 295 節) からなる新約聖書を用いた. 結果を表6に示す. この結果, 文章が長くなったことに伴なって, 全体的に一致度が下がっている. 一致度の中では, 周辺キーワードと中

*11 10 文以上の記事に限定している (平均 21.3 文).

表6　抽出する重要文と一致度の関係2（名詞・用言）

要約文の数	周中	周特	中特	周中特
1	0.519	0.148	0.407	0.148
2	0.574	0.167	0.333	0.167
3	0.617	0.222	0.383	0.210
4	0.657	0.306	0.454	0.287
5	0.685	0.321	0.494	0.309
10	0.719	0.419	0.585	0.404

心キーワードの一致度が最も高くなっている．これは，文章が長くなり頻度による影響が増加し，高頻度の中心キーワードとの間で相関が高くなったためと考えられる．

またいずれの例においても，(周辺，中心)（中心，特徴）の一致度が（周辺，特徴）よりも高くなっていることから，中心キーワードが周辺キーワードと特徴キーワードとをつなぐ役割を果たしており，本システムの意図する頻度情報からは得られないキーワードを中心キーワードという展望台からの眺めによって発見していると言える．

また表4から表6中の（中心，特徴）の一致度において，要約文数が1から2に増えた時に一致度が減少している．これは，高頻度の単語を最も含む文には，特に特徴のある単語が含まれやすいことを示す結果であり，この解釈としては，「文章には最も主張を表す明確な1文がある」ことを示す結果と言えよう．

4·2　キーワード抽出法の組合せの効果

一つのキーワード抽出法によって重要文抽出を行なうシステム，あるいはさまざまなキーワード抽出法を単純に組み合わせてその多数決によって重要文を抽出するシステムの場合，取り出される重要文には一貫性がないことが多い．しかし本システムの場合，文章の主題を表すキーワードを含みかつ特徴的なキーワードを含むように，意味のある絞り込みを行なえるキーワード抽出法の組合せを実現している．

この文章の主題は，文章の中で頻度の高い単語（周辺キーワード）と同時に出現する単語として，主に高頻度のキーワード集合としている．現在はこのキーワード集合を観点とした重要文を抽出しているが，中心キーワードとして別の観点に基づくキーワード集合を用意すれば，異なる観点に基づく重要文抽出にも応用が可能である．2·1の言葉を借りて換言すると，展望台を自分の好みの場所に設置して，辺りの特徴を見回すことが可能となるのである．

表7　先頭の文を含む確率（名詞）

要約文の数	周辺	中心	特徴	周中特	ランダム
1	0.463	0.400	0.362	0.318	0.047
2	0.575	0.517	0.476	0.423	0.094
3	0.645	0.607	0.570	0.500	0.141
4	0.699	0.662	0.633	0.565	0.188
5	0.743	0.714	0.690	0.628	0.235
6	0.779	0.757	0.738	0.679	0.282
7	0.811	0.793	0.778	0.725	0.329

表8　先頭の文を含む確率（名詞・用言）

要約文の数	周辺	中心	特徴	周中特	ランダム
1	0.449	0.342	0.325	0.261	0.047
2	0.580	0.478	0.430	0.365	0.094
3	0.653	0.564	0.497	0.434	0.141
4	0.708	0.621	0.533	0.474	0.188
5	0.753	0.675	0.560	0.509	0.235
6	0.789	0.721	0.578	0.537	0.282
7	0.819	0.758	0.591	0.556	0.329

4·3　冒頭文を含む確率

続いて，文章の冒頭文を含む確率について考察する．表7と表8とにニュース記事を用いた場合の確率値を示す．先頭の文を含む確率は周辺，中心，特徴の順に減少している．表8の最重要文についても，頻度によるキーワードのみでは先頭の文が44.9%の確率で取り出されるが，3種類のキーワードの組合せによると26.1%の確率となっている．これは，文章の中で最も重要な意味のある文は，冒頭文よりもむしろ文中の内容に含まれることを示す結果である．

5.　結論と今後の展望

文章中に含まれる文章の流れに沿いかつ特徴的なキーワードを見出す手法を提案し，3種類のキーワード抽出法の組合せによる重要文抽出法を提案した．このような文章の抄録を用いた応用はさまざまな可能性を秘めている．

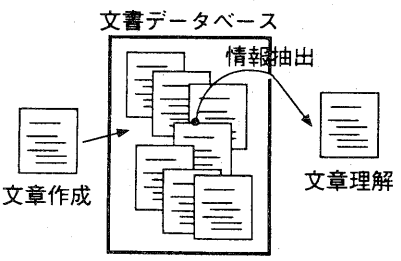

図4　情報伝達の流れ

我々は特に，文章そのものの作成と読解に関する支援システムを考えている．文章の持つ本来の役割で

ある情報の伝達のために，情報を送る側と受けとる側のそれぞれの支援に用いられると考えている．それらは情報を受けとるための文章読解支援と情報発信の支援としての，文章作成の支援である．現在，必要な情報を探すための支援は情報抽出などで数多くの研究がなされており，冒頭で述べたWebページ検索への応用も一つの課題である．しかし，文章そのものの作成および読解を支援するシステムはあまり見られない．

キーワード抽出や重要文抽出の技術が向上しようとも，文章の質が悪ければ正しく動作しない可能性がある．一貫性のない文章や，飾り言葉が多い文章などがそれである．もちろん，システムの改良を重ねることで，どのような文章からも重要文を抽出できるような一般的なシステムの構築を目指すのも研究の一つの方向ではある．しかし，システムばかりが向上して人間が堕落しても困るのであって，システムと人間の間の両者の歩み寄りによってわかり良い文章の作成と重要文の抽出の間の両者の妥協点が見い出されると筆者らは考えている．

◇参 考 文 献◇

[Brandow 95] R. Brandow, K. Mitze, and L.F. Rau: Automatic Condensation of Electronic Publications by Sentence Selection, Information Processing & Management, Vol.31, No. 5, pp. 675 – 685, (1995).

[Chasen] http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html.

[平尾 99] 平尾努, 北内啓, 木谷強: 単語重要度と語彙的結束性を利用したテキストセグメンテーション, 情報処理学会自然言語処理研究会資料, 99-NL-130, Vol.99, No.22, pp.41 – 48,(1999).

[Luhn 58] Luhn, H.P.: The automatic creation of literature abstracts, In IBM Journal for Research and Development, Vol.2, No.2, pp.59 – 165, (1958).

[Morris 91] Morris, J.J. and Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics, Vol.17, No.1, pp. 21 – 48, (1991).

[難波 99] 難波英嗣, 奥村学: 書き換えによる抄録の読みやすさの向上, 情報処理学会自然言語処理研究会資料, 99-NL-133, Vol.99, No.73, pp.53 – 60,(1999).

[任 98] 任福継, 定永靖史：統計情報と文章構造に基づく重要文の自動抽出, 情報処理学会技術研究報告 NL125, Vol.98, No.48, pp.71 – 78, (1998).

[大澤 99] 大澤幸生, ネルスベンソン, 谷内田正彦: KeyGraph:単語共起グラフの分割・統合によるキーワード抽出, 電子通信学会誌論文誌, Vol.J82-D-I, No.2, pp. 391 – 400, (1999).

[奥村 99] 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, pp.1 – 26, (1999).

[Salton 97] Salton, G. and Buckey, C.: "Term-Weighting Approaches in Automatic Text Retrieval", Readings in Information Retrieval, pp.323 – 328, (1997).

[砂山 99] 砂山渡, 野村勇治, 大澤幸生, 谷内田正彦：Webページ検索におけるユーザの興味表現支援システム, 電子情報通信学会論文誌, Vol.J82-D-I, No.12, (1999).

[Zechner 96] K. Zechner:Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, in proc. of 16th International Conference on Conputational Linguistics, Vol.2, pp.986 – 989, (1996).