データベース・システム 84−19
（１９９１．７．１８）

重要キーワード抽出方式とその活用方法

内山恵三，中村正規

東京電力（株） システム研究所

文書型データベースの検索キーワードを自動自動抽する際、文書から単純に言葉を切り出すため、１文書当りの抽出数が多く、人為的な洗練が必要となる。そこで、重要なキーワードが文書中の主語と目的語節に多く含まれることに着目し、不要語辞書を用いずに、文書の要旨に沿ったキーワードを抽出する方式を開発した。更に、その活用事例として連鎖検索を検討した。

DEVELOPMENT OF AN AUTOMATIC KEYWORD-EXTRACTING SYSTEM ON THE BASIS OF CONTENT ANALYSIS AND AN APPLICATION SYSTEM

Keizo Uchiyama , Masaki Nakamura

Computer & Communication Research Center

Tokyo Electric Power Company Co.

1-4-10,Irifune,Chuo-ku,Tokyo,104,Japan

An ordinary document database is a system in which the automatic extracting of keywords from documents is carried out indiscriminately; therefore the amount of keywords excised from each documents is too large,and it becomes necessary to control the amount in some way. Special attention is drawn to the fact that subjects or objective clauses contain more useful words for retrieval,and a system in which the extracting of keywords is implemented on the basis of content analysis has been developed without using any dictionary of useless words.and we have examined a retrieval system which have association keywords index.

1．はじめに

近年、情報ネットワークの発達により、各種情報のデータベース化が進み、その種類も社内データベースや商用データベースなど多種多用となっている。文献など文書型のデータベースでは、検索処理を迅速かつ効率的に行うために、検索キー情報として「キーワード」を付与している。

キーワード付与の方式は、古典的には、シソーラスなど人的付与の方式が採られていたが、最近は、コンピュータによる自動抽出処理が一般的である。この種のシステムは、キーワード付与作業の簡素化を狙いとしたもので、データベース構築の支援システムとして多大な貢献をしていることは間違いない。しかし、ハイパーメディアのような高度利用の支援を考えた場合まだまだ改良の余地があると考えられる。

本報告では、まず、キーワード抽出の現状の問題点を整理し、その解決策として我々が開発した重要キーワード抽出方式について述べ、最後にその活用方法を紹介する。

2．キーワード抽出の現状の問題点

実用化されているほとんどの抽出システムは、文書を単語単位に分かち書きし、キーワードに適さない不要語を除去した残りの言葉をキーワードにする方式を採っている。しかし、この方式は、文書の内容を考慮することなく、単純に言葉を切り出す為、文章のスタイルにもよるが、１，０００文字の文章から４００文字（１００ワード）程度のキーワードが抽出される。ＪＩＣＳＴや新聞データベースなど商用データベースでは、１文書に付与されているキーワードの数が３０ワード程度などで、これと比較すると機械抽出は非常に多くのキーワードが付与されといえる。このことは、検索時に検索漏れを無くす利点はあるが、反面、検索結果の中に検索意図と合わない情報が多く含まれていることを意味している。

抽出キーワードを洗練する方法として、

①人為的に重要キーワードを選別する

②発生頻度の高いものをキーワードにする

③不要語辞書を参照して削除する

が一般的に用いられているが、それぞれ以下の課題を持つ。

人為的に重要キーワードを選別する方式は、文章の内容に沿ったキーワードを得ることができるが、選別者の経験や理解度の影響を受けやすく、自動抽出の利点であるキーワードの均一化を保障することができない。

発生頻度の高いものをキーワードにする方式は、重要な言葉が、文書中で使用される頻度が多いという前提に基づき、抽出ワードの発生回数の累計を参照して重要度を決定する方式である。しかし、この方式は、新聞データベースのようにデータベースが扱う対象分野が広い場合には有効であるが、技術情報のデータベースのように対象分野が狭い場合は必ずしも有効の手段とはいえない。たとえば、研究報告のデータベースなどでは、「開発」、「評価」などの言葉が頻繁に使用されるが、これらはどちらかというとキーワードとして適さない言葉である。また、新造語など、発生回数の累計値が小さい言葉は、必ずしも正しく評価されない。

不要語辞書による方式は、利用者が作成した不要語辞書を参照して、キーワード候補から不要語を除去する方式である。しかし、不要語とは、文章の内容から判断すべきもので、言葉の要・不要を辞書で管理するのは最適な方法とはいえない。

今後、データベースが大規模化し、且つ、戦略的に活用していくためには、洗練したキーワードの抽出は不可欠なものである。

又、キーワードは、文書中の発生位置や文書中での重要度など潜在的な情報を多く含んでいる。例えば、文書中の発生位置は、画面に表示された文書中のキーワードを色表示する情報として、重要度は、文書間の最適リンクを張る情報として、ハイパーメディアのような高度情報検索システムには大切なものである。このような観点から、キーワードの他に、キーワードが持つ潜在情報を一緒に抽出する方式も重要な研究課題と言える。

3．重要キーワード抽出方式

3．1　基本的な考え方

文章の内容に沿ったキーワードを抽出するためには、本来、文章の意味を理解しなければならないが、そのための辞書は、国レベルで研究中の段階で、実用化までにはまだ時間を要すると思われる。

そこで、我々は、重要キーワードが文書中の主語と目的語の節に多く含まれていることに着目し、全文からキーワードを抽出するのではなく、これら重要文節からキーワード候補を抽出し、更に、候補ワード前後の格助詞などの情報から重要度を決定するアプローチを試みた。

3．2　重要キーワード抽出処理のステップ

重要キーワードは、表1に示すように、①分かち書き処理②文節の切出し③重要文節の抽出④キーワードの抽出⑤キーワードの重み付け⑥キーワードの絞り込みの6ステップから抽出される。各処理の役割りについて以下に述べる。

①．分かち書き処理

この処理は、文節単位に文章を切り出す処理を効率的に実行するための前処理で、文章を品詞単

表1　重要キーワード抽出処理のステップと適用ルール

ｽﾃｯﾌﾟ	処理	内容	適用ルール	主なルール
1	分かち書き処理	・対象テキストを単語単位に分かち書きする。		
2	文節の切り出し (重要文節の区切り)	・分かち書きされた文章の中から、重要文節の区切りとなる言葉を見つけ、文節として切断する。	文節切り出しルール	・文章切り出し ・読点による切断 ・主要格助詞による切断 ・強調キーワードの取り出し ・括弧内文節の対処 ・きめ細かい切断
3	重要文節の抽出	・切り出された文節の中で、主語、目的語を含む主要な文節（主語節、目的語節）する。	文節抽出ルール	・主要格助詞による抽出 ・不要文節削除
4	キーワードの抽出	・重要文節の中から、キーワードを抽出する。 ・また、文書のタイトル、及び強調された文節からキーワードを抽出する。	キーワード抽出ルール	・テキストキーワードの抽出 ・タイトルキーワードの抽出 ・強調キーワードの抽出
5	キーワードの重み付け (キーワードの重要度計算)	・キーワードについて発生文節の位置関係、発生頻度、言葉の属性などから、各キーワードの持つ重みを計算し、重要度の軽重を付ける。 ・最も重要度の高いものが[1.0]、逆に重要ではないと思われるものが[0.0]に近づくように、基本点の設定、及び加点と減点を行う。	キーワード重み付けルール	・文節出現位置による基本点付与 ・タイトルキーワードによる加点 ・接頭接尾文字による減点 ・頻度による重み付け ・強調キーワードによる加点
6	キーワードの絞り込み	・各キーワードの重要度を比較し、重要キーワードを絞り込む。	キーワードの絞り込みルール	・所定値以上による選択 ・抽出個数制限による絞り込み

位（単語単位）に分かち書きする。

②．文節の切出し

分かち書きされた文章の中から、文章の区切りとなる表記の特徴を見つけ、文節切り出し処理を行う。文章の区切りを示す最も明確な記号は句読点などで、まず、文章を句読点単位に分ける。句読点で分けた大文節は、主語や目的語の他に状況を説明するための副詞節や形容詞節などから構成されているため格助詞で更に細かく文節の切断処理を行う。

また、鉤括弧（「」、『』等）で括られた言葉を強調キーワードとして取り出す。これは、通常文書作成者が強調したい箇所（言葉）を鉤括弧で括ることが多いからである。

③．重要文節の抽出

切り出だされた文節の中から、「が」、「は」、「を」などの主語や目的語に使われる格助詞で終了している文節を重要文節とみなし、それ以外の文節を削除する。

④．キーワードの抽出

キーワードは、上記処理の結果得られた強調語及び重要文節と原文タイトルから抽出処理される。ここで、タイトルは文書の内容を端的に表している文であるが、タイトル中で使用している言葉が、本文中で出現しなかったり、重要文節から漏れる恐れもあるため、本文の記載に関係なく無条件に重要キーワードとして扱うよう考えた。

⑤．キーワードの重み付け

抽出されたキーワードは、このままでは、従来の抽出システムと同様に全て同レベルの扱いとなる。そこで、発生文節内でのキーワードの位置関係と強調語・タイトルキーワード（タイトルから抽出したキーワード）など言葉の属性から、各キーワードの重要度（基準値）を決定し（たとえば、抽出キーワードが、主語節を表す「が」や「は」の格助詞の前にある場合は、そのキーワードに高い基準値を与える）、文書中に何回も使用されている場合は加点する重み付けの計算を行い、接尾文字付きのキーワードは減点する重み付けの計算を行うなど、キーワードの重要度に軽重を付ける。

文書中に複数回キーワードが出現する場合の加点する計算式を以下に示す。

$$V=(V_0+V_1)-(V_0\times V_1) \quad \cdots (1)$$

但し、V_0は１回目に出現されたキーワードの基本点、V_1は２回目に出現されたキーワードの基本点であり、求められるＶがそのキーワードの重要度である。そして、同一のキーワードが３回以上出現される場合には、上記求められたＶを式（１）のV_0とし、新たに出現されたキーワードの基本点をV_1として算出し、以下それを所定数だけ繰り返し行えば良い。

接尾文字付きのキーワードの減点は、キーワードの基本点に一定の数値（（１以下の正の数）以下、「減点係数」と称す）を掛けることによって行う。この接尾文字としては例えば「名、量、風、策、図、表、化、系、圏、材、者、種、数、製、説、側、属、値、的、度、費、部、法、用、派、比、率、流、列、例、論、画、群、型、欄、点、性、日、時、類」等がある。しかし、上記のような接尾文字が付いているキーワードでも例えば「女性」、「温度」等のように重要キーワードとなり得るものもあり、一様に同一減点対象とするのはまずい。そこで、接尾文字を含めて２文字の場合は減点を少なくし、重要度が高くなるようにした。具体的には、２文字の場合、

$$V'=V_0\times 0.7$$

その他の場合、

$$V'=V_0\times 0.5$$

とした（但し、V_0は各キーワードのもつ基本点である）。

⑥．キーワードの絞ぼり込み

最後に、抽出された重要キーワード候補の中からデータベースに登録するキーワードを決定する処理で、抽出個数制限と重要度の評価値制限による絞ぼり込み方式の中から、利用者が、アプリケーションによって選択する。

４．重要キーワード抽出の実施例

前述した抽出システムの実施例について図１を基に説明する。まず、対象テキストの本文に対し

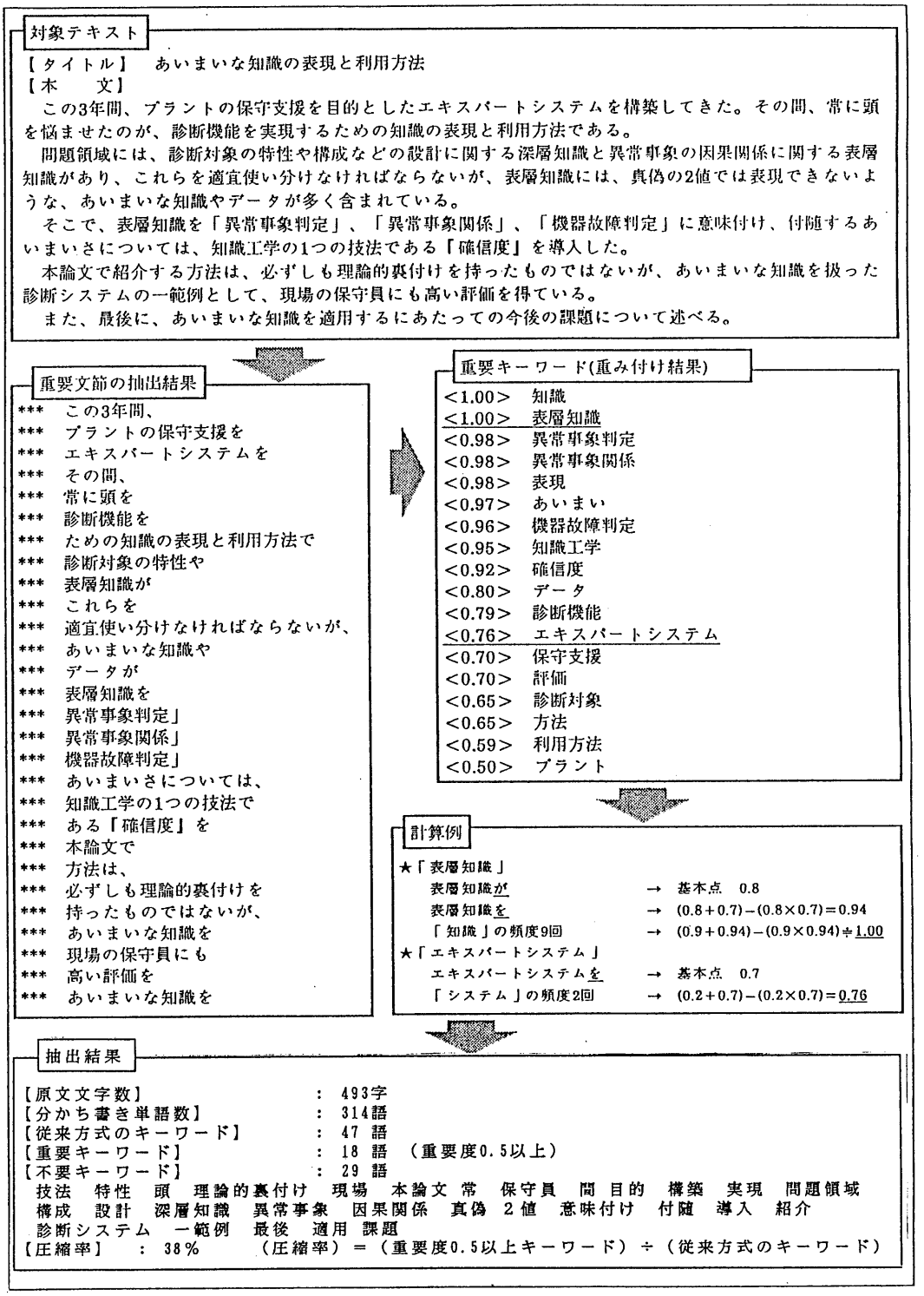

図1. 重要キーワード抽出システム実施例

て分かち書き処理を行い、文節切断ルールにより文章を格助詞などで詳細に切り出す。図1の「重要文節の抽出結果結果」は、詳細に切断された文節の中から、主語や目的語を含む重要な文節を抽出したものである。「表層知識が」や「プラントの保守支援を」など、文書中の主語節や目的語節が抽出される。

次に、抽出された重要文節からキーワードを抽出し、重み付けを行う。例えば、「表層知識」というキーワードの場合、2つの重要文節に出現している。まず、初めの文節では、格助詞「が」から、主格として用いられていると判断して基準点0．8を与える。2つ目の文節では、「を」を基に目的格と判断し、基準点0．7を与え、両者の基準点から重み付けを計算し、0．94を得る。さらに、最短単位のキーワード「知識」の頻度（9回）により、重み付けが行われ、「表層知識」の重要度は1．0と計算される。

他に、タイトルキーワード、強調キーワードによる加点、あるいは、接頭・接尾辞や言葉の属性などによる減点などの重み付けが、適宜適用される。なお、重要度は、0．0から1．0の値をとり、1．0に近いほど、文書の内容を示す重要なキーワードであることを示す。

本実施例では、重要度が0．5以上のものを重要キーワードとするように絞ぼり込みを行っている。その結果、本テキストにおける重要キーワードは18語となり、従来方式のキーワード自動抽出の語数（47語）と比較すると、圧縮率は38％となる。

5．重要キーワード抽出システムの評価

5．1　キーワードの削減効果

文書は、対象分野や目的により、そのスタイルが異なってくる。そこで、対象分野が狭く、特有のスタイルを持つ研究報告書、対象分野は狭いが、読み手の層によって、その表現に幅がある技術論文、対象分野が広く表現が一般的である新聞記事、これらのおのおのについてキーワードの削減効果を分析した。

表2　重要キーワード抽出システムによるキーワード圧縮率の効果（対象分野別）

項　　目	研究報告DB	技術論文DB	新聞記事DB
文字数	565字	439字	857字
分かち書き単語数	271語	261語	427語
従来方式によるKW数	64語	48語	96語
重要KW数 （重要度0．5以上）	43語［67．0］	25語［51，4］	60語［62．3］
重要KW数 （重要度0．6以上）	31語［48．0］	19語［38．8］	44語［45．9］

注：［ ］内は、圧縮率…従来のシステムのキーワード数を100とした時の重要キーワード数

［ ］以外の数は、平均値

その結果、表2に示すように、文書の種類によらず、キーワードの削減率（圧縮率）が一律に高まるとともに、重要度の分散効果も現われ、各キーワードの重要度が適切な値となり、重要キーワードとしてのノイズや漏れも減少し、質的に見ても満足できる結果が得られた。

なお、サンプルデータでは、重要度が0．6以上のキーワードを実質的な重要キーワードとすることが可能となり、抽出キーワード数において、従来方式の50%削減を実現した。

5．2　キーワード抽出方式の評価

次に、情報処理分野における技術論文を基に、さらに詳細な分析を行った。

本分析に使用した論文の抄録の平均文字数は、約400文字である。また、文書中で使われている助詞、名詞、形容詞、副詞などの単語数は、約200語（重複を含む）で、そのうち、名詞（キーワードと成りうる単語、これを1語基と呼ぶ）は約50語である。実際のキーワードは、この1語基か、あるいは、1語基を組み合わせた言葉である。例えば、キーワード「表層知識」は、1語基の「表層」と「知識」を組み合わせたもので、これを2語基と呼ぶ。従来方式によるキーワード抽出では、過去の経験から、最大3語基（3つの1語基の組み合わせ）まで認めている。

また、重要キーワード抽出処理における絞り込みの基準は、文章の形態や筆者の文章の書き方のくせを加味し、適切な評価を行うために、0．5以上を基準にして検討をした。

まず、キーワード抽出数についてみると、表3に示すように、従来方式では平均65．9語のキーワード数に対し、重要キーワード数は平均20．4語と大幅に削減していることが分かる。その結果、従来方式により抽出されたキーワード数を100語とした場合の重要キーワード数は、平均31語（圧縮率31%）となる。

なお、圧縮率が高いもの（20%未満）、及び低いもの（55%以上）について、その内容を分析した。その結果、このような圧縮率になる原因としては、文書テキストが極端に長い（文字数が多い）場合や、逆に短い場合、あるいは、システム名や機能名などにより、キーワードの組み合わせパターンが多くなる場合であることが分かった。そこで、文字数、及び1語基キーワードに対する抽出度合により抽出結果を正規化した。その結果、

表3－1　キーワード抽出数

項　目	最　大	最　小	平　均	標準偏差
文字数	761字	181字	390．2字	91．4
分かち書き単語数	447語	94語	221．5語	57．9
1語基KW数	102語	19語	47．2語	11．0
従来方式によるKW数	128語	25語	65．9語	17．3
重要KW数（重要度0．5以上）	48語	8語	20．4語	6．6

注：1語基KW…自動抽出システムより抽出された単語（組み合わせなし）

表３－２　圧縮率（従来方式により抽出されたキーワード数を１００とした時のキーワード数）

項　　目	最　高	最　低	平　均
圧縮率 （重要ＫＷ数／従来方式のＫＷ数）	１２％	６７％	３１％

表３－３　文字数、１語基ＫＷに対する抽出度合

項　目	最　大	最　小	平　均	標準偏差
従来方式のＫＷ数／文字数	０．２９	０．０９	０．１７	０．０４
重要ＫＷ数／文字数	０．１０	０．０３	０．０５	０．０１
従来方式のＫＷ数／１語基ＫＷ数	１．９０	１．０３	１．３９	０．１６
重要ＫＷ数／１語基ＫＷ数	０．６９	０．１８	０．４３	０．０９

（注：重要ＫＷ数は重要度０．５以上を対象とする）

文字数に対しては、従来方式では１７％、重要キーワードでは５％の割合でキーワードが抽出され、また１語基に対しては、おのおの１３９％、４３％の割合で抽出されることが分かった。

これらの評価グラフを図２、図３に示す。グラ

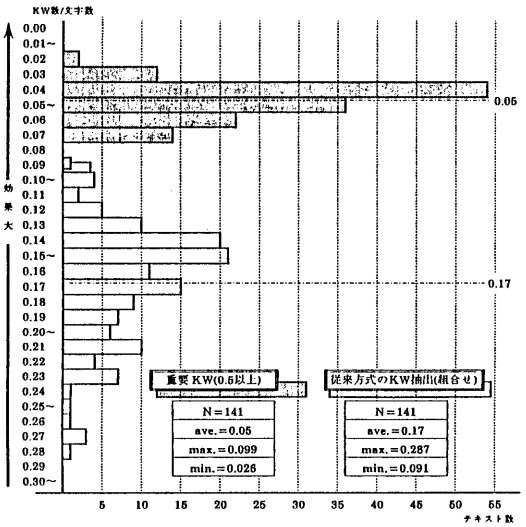

図２　文字数当りのキーワード抽出数

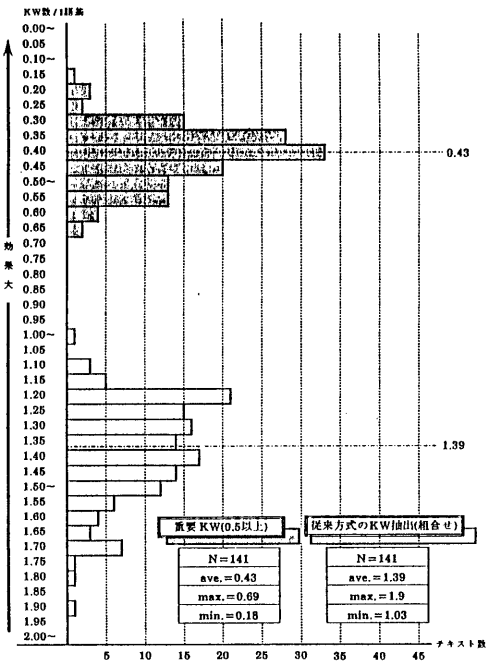

図３　キーワード最短単位（１語基）に対するキーワード抽出数

フから、従来方式に比べて、重要キーワードによる抽出結果の方が、分布域の幅が狭く、文書の文字数（長さ）や用いられている１語基の記述形式などの影響を受けにくく、均質なキーワード抽出処理が行われていることが読み取れる。

５．３　評価のまとめ

前述した重要キーワード抽出のシステム化及び評価を通して、厳密な意味解析を行わなくても、簡単な抽出ロジックとルールで重要度の高いキーワードを抽出できることが確認できた。その結果、文書の内容を表していない言葉や、たまたま表れた言葉がキーワードとなることを抑制し、平均して、従来方式の５０％削減を実現した。

また、キーワードとその重要度から「文書の重要度」、つまり、主題に対する文書の適合性が判断できる。これにより、検索者の意図に従った検索が行え、欲しい情報が的確に入手できるようになる。

なお、キーワード抽出過程の副産物である重要文節を組み合わせることにより、文書の概要を把握することもできる。

５．４　適用領域と応用分野

本システムは、キーワードの抽出機能に加えて、文節の切断機能、言葉の重み付け機能を持つ。文章に使われている言葉に重要度が付くということは、自然言語処理において非常に有用な機能であり、様々な分野への応用が考えられる。

自動索引技術はもとより、文章を自動的に区分けする自動分類技術、文章中の必要かつ重要な文を取り出す文章抽出・加工技術、あるいは文章解析技術などへ「重要度」の概念を適用することが可能となる。これにより、要旨抽出・抄録作成、主題分析、自動タイトル付け、シソーラス（言葉の辞書）作成、文章整形や文章校閲、あるいは自由記入のアンケート評価・分析、特許情報をはじめとする情報分析、さらにはファジィデータベースなど、より知的なシステムの実現が図れる。

６．連鎖検索への活用

６．１　文書データ間の関係付け

一般にデータベースの検索を行う場合、対象となる内容の詳しい説明が知りたい、対象となる内容に関連したデータが知りたいなどの検索要求に合せて検索行為が進められる。しかし、これらの行為を実現するためには、予めデータベース内で対象となるデータと検索データ間を結び付けるという作業を施す必要がある。現状、このようなデータ間の関係付けは、ハイパーメディアやハイパーテキストで代表されるように手作業で行われている。

文書データ間の関係付けは、①文書データの内容による関係付け（内容が類似している、反対の内容である、やさしく書かれてある、詳しく書かれてあるなど）、②文書データ中の言葉による関係付け（同義語、類義語、関連語など）に大別される。前者の場合、この関係付けを自動化するためには、人間が関係付けする時に無意識に用いている判断基準や知識をルール化あるいは辞書化する必要があり、多くの人が研究を進めている。今回提案した重要キーワード抽出方式は、ある程度文章の内容に沿ったキーワードを抽出できる点を利用して特殊な辞書無しで文書データ間の関係付けが適切に張れるかどうか検討してみた。

６．２　関係付け処理のステップ

文書データ中の言葉と他の文書データの関係付け表は、以下のステップを経て作成する。

①．タイトル別キーワードの抽出

文書ごとに重要キーワードを抽出し、タイトル別のキーワード・重要度の関係を抽出する。

②．キーワード別タイトルの関係に変換

キーワード別にタイトルの並び替えを行ない、他の文書データと関連がないキーワードを取り除く。

③．文書データの重要度による並び替え

同一キーワードを持った文書データを重要度の高い順に並び替えることによって、キーワードに対する文書データの順位付けを行う。

6．3　関連付け処理結果例

［タイトル：高速増殖炉］

キーワード	重要度
高速増殖炉	0.99
常陽	0.95
もんじゅ	0.90
開発	0.70
タンク型炉	0.70

キーワード	重要度
設計研究	0.70
21世紀	0.50
ウラン	0.50
前半	0.50
原型炉	0.50

キーワード	重要度
原子炉	0.50
ループ型炉	0.50
軽水炉	0.50
実験炉	0.50

［タイトル：軽水炉でのプルトニウム利用］

キーワード	重要度
プルトニウム	0.99
軽水炉	0.98
燃料	0.97

キーワード	重要度
常陽	0.95
ウラン燃料	0.93
ウラン	0.91

キーワード	重要度
プルトニウム利用	0.74
原子炉	0.70
高速増殖炉	0.50

［タイトル：再処理における脱被服技術］

キーワード	重要度
脱被覆技術	0.95
プルトニウム	0.70
再処理	0.70

キーワード	重要度
有機溶媒	0.70
被覆管	0.70
ウラン	0.50

キーワード	重要度
溶解	0.50

重要キーワード抽出システムによって上表のように重要キーワードが抽出されたタイトル「高速増殖炉」「軽水炉でのプルトニウム利用」「再処理における脱被服技術」の３つの文書データ間の関係付け結果を次表に示す。

キーワード	タイトル	重要度
高速増殖炉	・高速増殖炉	0.99
	・軽水炉でのプルトニウムの利用	0.50
常陽	・高速増殖炉	0.95
	・軽水炉でのプルトニウムの利用	0.95
ウラン	・軽水炉でのプルトニウムの利用	0.91
	・高速増殖炉	0.50
	・再処理における脱被覆技術	0.50

キーワード	タイトル	重要度
原子炉	・高速増殖炉	0.50
	・軽水炉でのプルトニウムの利用	0.50
軽水炉	・軽水炉でのプルトニウムの利用	0.98
	・高速増殖炉	0.50
プルトニウム	・軽水炉でのプルトニウムの利用	0.99
	・再処理における脱被覆技術	0.91

キーワードは文書データのタイトル毎に重要度が異なっている。重要度の高いキーワードを持つ文書データほどキーワードの関連が強いため、文書データを検索する場合、同一キーワードを持つ文書データのうち重要度の低い文書から高い文書へ検索する方が有効的である。例えば、高速増殖炉の内容を知りたいという検索者に対して、「高速増殖炉」というキーワードで検索を行う場合、タイトル「軽水炉でのプルトニウム利用」の文書データ中の「高速増殖炉」という言葉からタイトル「高速増殖炉」の文書データの検索は有効であるが、その逆の検索は有効ではない。

このように、重要キーワードの重要度によって文書データ間の関係付けの方向性がある程度判別できる。

6．4　関係付け処理の評価

１文書当りの平均２３７文字の技術文書４９件について、一般によく使われている、文書データから単純に言葉を切り出すキーワード抽出方式を用いて文書データの関係付け処理を行った場合と本方式で処理した場合の結果の比較を次表に示す。

従来方式との比較表

	従来方式	本方式
１文書の平均キーワード数	２８	１４
１文書にリンクする平均キーワード数	２８	７
キーワード・文書間のリンク付け総数	３３１	１０４
１キーワードにリンクする平均文書数	４	３
１キーワードにリンクする最大文書数	３６	１６

従来方式では、文書中に他の文書のキーワードが存在すれば、その文書が他の文書のキーワードとの関連性が薄くてもキーワードとしてしまうため、文書中のキーワード全てに他文書とのリンクが張られている。それに対し本方式は、文書中のキーワードと関連性の強い多文書だけにリンクを張るためリンクキーワード数も１４キーワード中７個

と少なく洗練されていることが分かる。また、１キーワードにリンクする文書数が同程度なのに対し、キーワードと文書間のリンク付けの総数は従来方式の３分の１に本方式は減少している。このことからも無駄なリンク付けが解消されたことが分かる。

以上のことから、連鎖検索する際の文書データ間の関係付け処理として本方式が有効であると考えられる。

７．今後の課題

以上述べてきたように、重要キーワード抽出処理は、おおむね良好の結果が得られたが、今後の課題として次の２点があげられる。

文節の切断や重要度の評価は、文書中の句点や助詞など表記の特徴に着目し、それらが持つ用法までの解析は行っていない。句点や助詞は文章の構造によってその用法が違う為、例えば、「重要度が」のように主格を示す［が］と「候補としてあげていたが」のように意味の上で次の文節の条件を示す［が］では、その文節の重要性に差があり、この例では、後者の文節は、不要な文節といえる。現行の方式では、両者とも重要文節として判定しキーワード候補として「候補」も抽出される。最終的には、重み付け処理の時点で「候補」の重要度は低く評価されるが、抽出結果の精度を高めるために、用法の解析ルールの検討が必要であると考えている。

もう１つの課題としては、キーワードの抽象化があげられる。人間がキーワードを付与する場合、本文中の内容を端的に表したいために、本文中で使われている言葉の他に、いくつかの言葉をまとめて抽象化したキーワードを付与することもある。このような原文中にないキーワードを扱うためにはキーワードの意味モデルが必要となるが、これは検索者の利用者モデルに大きく依存するものである。

このような課題の解決により、重要キーワードの活用事例であるデータベースの連鎖検索のための文書データ間の関係付け処理もより実用性の高いものになると考えられる。

連鎖検索の今後の課題としては、以下の２点があげられる。

文書データからキーワードにより他文書を検索する場合、キーワードに関連する文書がたくさんあると、どの文書が検索者が一番必要としている文書なのかという順序付けが必要になってくる。重要キーワード抽出方式の活用事例ではキーワードの文書における重要度によって検索文書の順位付けを行ったが、広い分野で使われるキーワードが分野別に高い重要性を持っている場合の的確なキーワードの順位付け、検索文書を検索するまでの到達時間の短縮のためには一層の検討が必要である。

もう１つの課題は、文書中のキーワードと同じように使われたり、同じ意味を持った言葉が他文書で重要性が高くなって存在する場合のその文書との関係付け処理である。文書中のキーワードと他文書を関係付ける時、そのキーワードの同義語や類義語を含めて検索対象とすることにより検索漏れが少なくなる。

８．おわりに

今回提案した重要キーワード抽出方式は、一般的に、ロール付きキーワード方式といわれるもので、キーワードが潜在的に持つ情報の１つとして重要度という指標をキーワードから自動的に抽出し、その役割りに幅を持たせたものであるが、研究としてはまだ出発点と言える。また、本方式を活用した連鎖検索も開発の初期段階である。今後、キーワードが持つ潜在情報を体系的に整理・具現化し、知的な検索処理の支援に寄与できたら幸いである。