1G-5

係り受け関係を用いた類似文書検索システム

稲垣　博人，宮原　末治，加納　英文，小橋　史彦
NTTヒューマンインタフェース研究所

1．はじめに

現在、大量の文書が電子化されつつあり、その電子化された大量の文書の中から、必要な情報を迅速かつ的確に収集する情報スキャニング技術が必要とされて来ている。従来の検索システムでは、入力に検索用の論理式や自然言語入力［1，2］等を用いるのが一般的であった。しかし、ある文書に類似した文書を検索しようとする場合、従来の手法では、膨大な論理式が必要となったり、文書のテーマを把握しなければならないなどの問題点がある。

本システムは、文書を直接認識入力し、その入力文書と最も類似した文書をキーワードや係り受け関係を利用して検索するシステムである。

2．類似文書検索システムの概要

本システムは、文書の入力･蓄積･検索という一連のシーケンスに即した処理を可能とし、漏れの少ない検索を実現することを目標とする。本システムの構成を図１に示す。

文書入力部は、主に文字認識部とイメージリーダから成る。［3］

文書加工･検索部では、入力された文書の構文解析及び索引抽出処理を行う文書加工部と、入力文書に類似した文書をデータベース中から抽出する類似文書検索部とからなる。文書加工･検索部の処理フローを図２に示す。

文書データベース管理部では、加工した文書を蓄積すると同時に、検索部で即利用できるように検索情報の更新を実施する。検索対象には、類似文書検索の必要性が高い特許文書を使用した。

3．文書加工部

（1）　形態素解析と係り受け解析

形態素解析では、入力文書を単語単位、文節単位に分割すると共に、品詞情報、意味カテゴリ情報等を付与する。

係り受け解析では、文節の係りと受けの関係を一意に決定する。日本語では、係り受け関係の曖昧性が高いので、文書中に出現した係りと受けの関係を記述する意味連結パターンを曖昧な係り受け関係に適用することにより曖昧性を減少させている。［4］

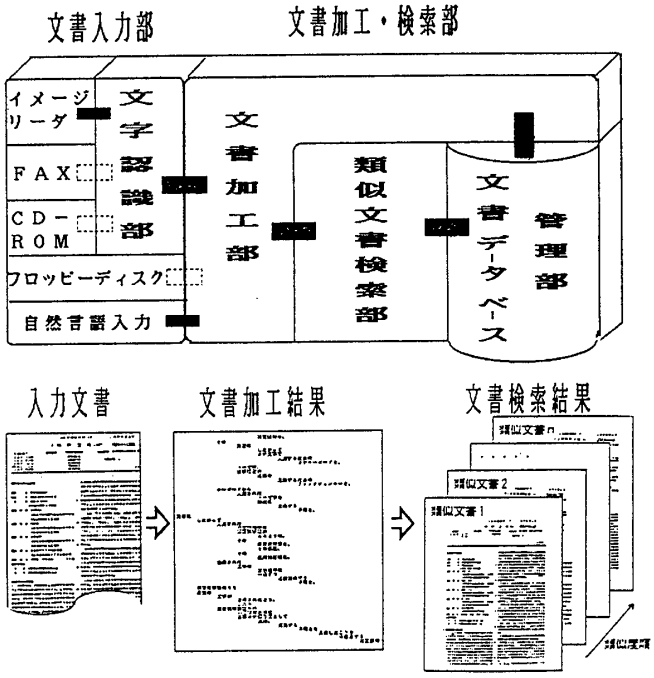

図１　類似文書検索システム構成図

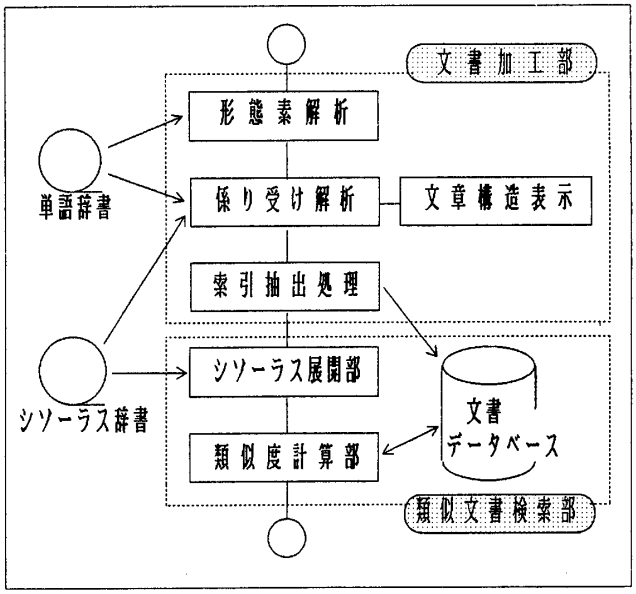

図２　文書加工･検索部の処理フロー

Document Retrieval Based on Modification Relations
Hirohito INAGAKI, Sueharu MIYAHARA, Hidefumi KANOU, Fumihiko OBASHI
NTT Human Interface Laboratories

（２）　索引抽出処理

索引抽出処理においては、文書中に出現している名詞相当語句から不要語を除いた単語をキーワードとして抽出する。

不要語辞書としては、自立語不要語辞書と接辞不要語辞書を用意した。

	例
自立語不要語	場合,特徴,一方 ,該,前記...
接辞不要語	計算機間の...，各車両は...

４．類似文書検索部

（１）　シソーラス展開部

漏れの少ない検索を実現するために、シソーラス辞書でキーワードを同義語及び類義語に展開している。その際、同義語及び類義語を意味の類似性が高い順に、同義語０，同義語１，同義語２，類義語の４段階に分け、意味の類似性の高さに応じて重み付けした。［５］

（２）　類似度計算部

本システムは、シソラース辞書を用いて検索を行っているため、漏れが少なくなる反面、膨大な数の類似文書候補が出力される。そこで、出力された候補を文書の類似度の高い順に並べ替えを行っている。

具体的には、文書間の類似度は、以下の３つの処理から決定した。

①キーワード包含率検査
②類似係り受け関係処理
③単語の構文的、意味的重み付け処理

処理の流れとしては、①のキーワード包含率を検査したのち、キーワード包含率の高い候補に対して、②および③の処理を行った。

①のキーワード包含率検査では、２文書間のキーワードの一致個数をカウントし、一致個数が高い文書ほど類似度が高いとする処理である。但し、同一のキーワードが複数存在する場合、キーワードは一個としてカウントした。

②は、格関係、連体修飾関係、「の」の関係等の係り受け関係の中で、類似した係り受け関係が存在する場合、２文書間の類似度をさらに上げる処理である。この時、各キーワードの構文的及び意味的重要度を考慮した。

構文的重み付けとしては、係り受け解析より抽出した文章構造を基に重み付けした。特に、特許文のように文書構造が比較的明確な場合、その構造に準じた形で重要度を疑似的に付与することが可能となる。

意味的重み付けは、前述した同義語、類義語の意味的類似度に基づいた。

５．検索実験例

現在、データベース中には、特許請求範囲文約１５０件分（自然言語処理関連、特にかな漢字変換分野）のデータが蓄積されている。

検索例として、かな漢字変換装置の同音異義語選択に関する特許を入力文書とした場合を示す。

かな漢字変換に関する特許は、１５０件中に約５０件あるが、そのうち、入力文書に最も類似した特許は、５件あった。この５件が検索により類似文書候補として上位１０位以内に出現する様子を示したのが表１である。キーワードをシソーラス展開しない場合とシソーラス展開した場合とで比較してみると、シソーラス展開したキーワードによるキーワード包含率検査のほうが効果的であることがわかる。しかし、ただ単にキーワードの包含率を検査しただけであると、上位１０位までに希望する文書の４割（２件）しか得られない。しかも、かなり下位に出現している。

一方、キーワードの重み付けと類似係り受け関係処理を行った場合、希望する文書の８割（４件）が上位１０位以内に出現するという結果が得られた。このように、キーワードの重み付けと類似な係り受け関係の抽出が文書間の類似度を検査する場合、有効なパラメータであることがわかった。

表１　類似特許出現順位の比較（単位　件）

	1～3位	～5位	～10位
ｷｰﾜｰﾄﾞ包含率検査（ｼｿｰﾗｽ展開なし）			2
ｷｰﾜｰﾄﾞ包含率検査（ｼｿｰﾗｽ展開有り）		1	2
類似係り受け関係処理	1	2	4

６．おわりに

本稿では、係り受け関係を用いた類似文書の検索システムについて述べた。特に、文書間の類似度を表わす尺度として、係り受け関係に着目する処理を提案した。

今後の予定として、現在用いている類似度計算方法を改善し、さらに、適合率の高い類似文書検索方法を確立する。

謝辞　本研究を行うにあたり、特許データベース構築等で多大な協力を頂いたＮＴＴ技術移転㈱の清末三恵子嬢、木村淳子嬢に感謝致します。

参考文献

［１］杉山ほか：自然言語理解に基づく情報検索システムIRIS,情処学会研究会報告ＮＬ58-8,1986.
［２］福永ほか：語の類義性と結合関係を考慮したテキスト検索,信学会春季全国大会，D-304,1989.
［３］宮原ほか：文書情報蓄積検索システムの検討，情処第３９回全国大会,1989.
［４］稲垣ほか：意味連結パターンを用いた係り受け解析,情処学会研究会報告ＮＬ67-5,1988.
［５］加納ほか：情報検索用シソーラスの試み,情処第３９回全国大会,1989.