2N-2

ファジィ文書検索システム（1）
～実験システムと評価～

森田哲也　小川泰嗣　小林清彦
(株)リコー　中央研究所

1．はじめに

光ディスクファイリングシステムの普及により大容量データベースがオフィス・個人規模で利用可能となっている．しかしこれらのシステムは，ファイリング時には大量文書に対する分類やキーワード付け作業が煩雑であり，検索時には利用者の要求主題に適した文書がなかなか探し出せない等の問題点があった．これらを解決するためのファイリング支援方式として，自動分類方式・キーワード自動抽出方式の研究が，また検索支援方式としてファジィシソーラス[1]・キーワード関係知識ベース[2,3]等を用いた情報検索の研究が行なわれている．

我々は，[2,3]で提案されたキーワードコネクションマトリックスを用いたファジィ文書検索および学習方式を拡張し，また文書数10,000件についてキーワード自動抽出を行なうことによりキーワード数約2,000・コネクション数約72,000の知識ベースを持つ知的ファイリングシステムを開発した．本論文では，その構成および性能評価結果について述べる．

2．キーワードコネクションマトリックス

[2,3]では語彙関係を用いた概念空間として，キーワードコネクションマトリックス（以下，ＫＣＭ）を提案した．これは任意の２キーワード間の概念的類似度を定量的に表現したキーワード関連度を，知識として持つキーワード関係知識ベースである．ＫＣＭは，キーワード数をＫとするとき，Ｋ×Ｋの行列Ｗとして与えられる．キーワードｉとｊの関連度の初期値Ｗijは，

$$Wij = \begin{cases} \dfrac{Nij}{Ni+Nj-Nij} & : i \neq j \\ 1.0 & : i = j \end{cases} \quad \cdots (1)$$

Ｎi，Ｎj:キーワードｉ，ｊそれぞれの出現頻度
Ｎij　:キーワードｉとｊの共出現頻度

ＫＣＭは一種のシソーラスであり，その初期値は[1]におけるＲＴ(Related Term)の計算式と等価であるが，後述の学習によりキーワード関連度が変化する点で異なっている．

Fuzzy Document Retrieval System.(1)
--Experimental System and Results--
Tetsuya MORITA, Yasushi OGAWA, Kiyohiko KOBAYASHI
Ricoh,Co.Ltd. E-mail:morita@takezo.rdc.ricoh.junet

3．機能および処理概要

①自動ファイリング機能：

まず，入力文書を形態素解析した結果に語接続規則が適用され，キーワード候補が抽出される．候補から不用語を除去した後，既存キーワードとの比較・出現頻度処理が施され，結果がユーザに提示される．ユーザは文書内容に適したキーワードを抽出結果から選択し文書に付与することができる．

②あいまい連想検索（ファジィ文書検索）機能：

一般にキーワードと論理演算子(AND/OR/NOT)からなる検索式Queryは(2)式のような積標準形に変形できるため，ファジィ文書検索における結果集合は，検索式中の任意のキーワードＫに対して各文書が[0,1]の値をメンバーシップ値としてもつファジィ集合Ｄ(Ｋ)を用いて(4)(5)式のように表現される．即ち積標準形化されたQuery に対するファジィ検索の結果集合は，それぞれの副検索式SubQueryに対する結果を用いて表せる．

ここで，ＮおよびｈはSubQueryの総数および番号，ｎおよびｍはそれぞれ否定演算子（¬）のないキーワードと否定演算子付きのキーワードの数を表す．

$$Query=SubQuery(1)\wedge\ldots\wedge SubQuery(N) \quad \cdots(2)$$

$$SubQuery(h)=K_1\vee..\vee K_{nh}\vee\neg K_{nh+1}\vee..\vee\neg K_{mh} \quad \cdots(3)$$

$$Result=SubResult(1)\cap\ldots\cap SubResult(n) \quad \cdots(4)$$

$$SubResult(h)=D(K_1)\cup..\cup D(K_{nh}) \cup D(K_{nh+1})\cup..\cup D(K_{mh}) \quad \cdots(5)$$

集合演算として代数和・代数積を使用すると，副結果集合SubResult(h)の文書ｉのメンバーシップ値ｒi(h)は，

$$ri(h)=1-\Big(\prod_{Kj\in Q(h)+} Sij\Big)\Big(\prod_{Kj\in Q(h)-} Rij\Big) \quad \cdots(6)$$

where, $Rij=1-\prod_{Kk\in Ai}(1-Wjk)$, $Sij=1-Rij$

Q(h)+,Q(h)-:否定なし，付きの検索キー集合
Ai:文書ｉ中に含まれるキー集合

各文書確度ｒiは，全ての副検索結果の結合であり，

$$ri=\prod_{h=1}^{N} ri(h) \quad \cdots(7)$$

③適応化学習機能

ユーザは検索結果に対する適切さの評価値を入力することにより，キーワードコネクション中のキーワード関連度を変更できる．これを学習機能と呼び，検索結果の文書確度と適切さの評価値によって定義される評価関数に最急降下法を適用することによって実現した[4]．

4．システム構成

本システムでは，文書画像を取り込み文字コード化するＯＣＲユニットおよび文書・ＫＣＭ等を保持する光ディスクユニットが，Ｕｎｉｘ　ワークステーションに接続されている（Fig.1）．またソフトウエアはＣ言語で記述し，文書ファイリング支援を行なうキーワード自動抽出部と，あいまい連想検索および学習を行なうファジィ検索・学習部の２ブロックからなる．ユーザＩ／Ｆは Ｘウィンドウバージョン10を用いて開発した．

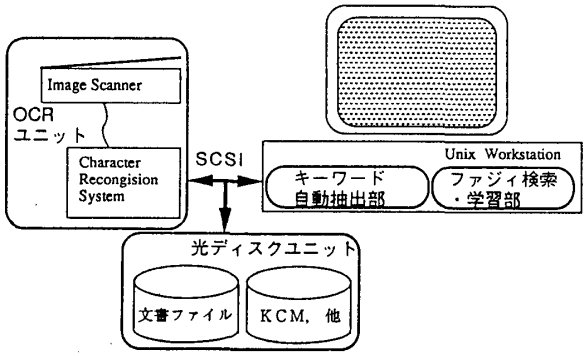

Figure.1　実験システムの構成

4．実験方法

技術文書10,000件からキーワード自動抽出機能により約2,000 キーワードを抽出，またこれらを基に72,000リンクからなるキーワードコネクションを作成した．

4．1　検索機能の性能測定方法

検索条件式として４つのキーワードードの単体および論理和／論理積／否定の組合せによって構成した．条件式に対して４人の実験者が個々に正解と見做した文書集合を，統計処理して正解集合を作成し，ファジィ検索の結果集合との関係を再現率・適合率で測定した．同様の検索条件に対する従来のクリスプな検索方式を用いた場合の再現率・適合率を測定し，上記結果と比較した．

4．2　適応化学習機能の性能測定方法

キーワード単体および論理和を用いた検索式の２種類について学習前の再現率・適合率と，学習を経過した後のそれらを測定した．学習は実験者によって作成された正解集合を適切さの評価値=1.0（正解）として30回の学習を試行した．

5．結果および評価

本ファジィ文書検索方式と従来方式について３種類の検索条件に対する性能比較をTable.1 に示す．

検索条件	ファジィ文書検索		従来方式	
	再現率	適合率	再現率	適合率
2ｷｰ論理積	52 %	5 %	22 %	6 %
2ｷｰ否定	72 %	19 %	23 %	7 %
1ｷｰ単体	66 %	40 %	44 %	62 %
平均	63 %	21 %	30 %	25 %

Table.1　各検索条件における性能比較

検索式として｛ＣＡＤ　ｏｒ　ＬＳＩ｝を用いた時の再現率(Recall)・適合率(Precision) vs. 閾値の関係をFig. 2に示す．閾値は検索結果集合を決定する文書確度の閾値である．また，検索式｛ＣＡＤ｝で正解集合を用いて学習を繰り返した結果をFig. 3に示す．

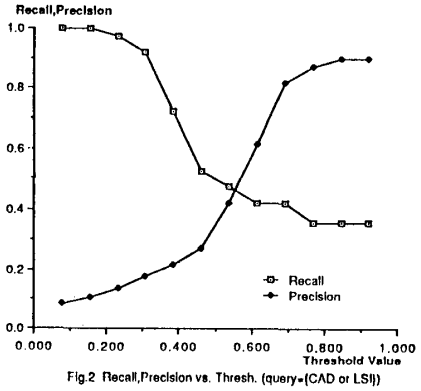

Fig.2 Recall,Precision vs. Thresh. (query={CAD or LSI})

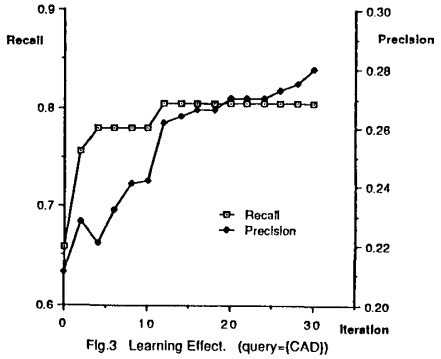

Fig.3 Learning Effect. (query={CAD})

6．まとめ

キーワード自動抽出・連想検索・適応化学習機能を持つファジィ文書検索システムをＷＳ上に構築し，検索および学習機能について再現率・適合率を測定することにより評価を行なった．従来のクリスプ検索に比べ連想による再現率の向上・および学習による効果が認められた．

7．参考文献

[1]S.Miyamoto, T.Miyake, and K.Nakayama.
"Generation of a Pseudothesaurus for Information Retrieval Based on Co-occurences and Fuzzy Set Operations". IEEE Trans. of SMC, vol.SMC-13, No.1, pp.62-70,1983.

[2]森田，小川，小林．
"動的キーワードコネクション方式を用いた文書検索システム"．昭63信学全大．vol.D1, pp.132, 1988.

[3]Y.Ogawa, T.Morita, and K.Kobayashi.
"Fuzzy Document Retrieval System and Its Learning Method Based on the Keyword Connection". Proc.of Int.Workshop on Fuzzy System Applications, pp. 143-144, 1988.

[4]小川，森田，小林．
"ファジィ文書検索システム（２）　～キーワードコネクションマトリックスの学習方式～"
情報処理学会第39回全国大会予稿集，1989.