内容検索システムとしてのＩＲＩＳ

杉山健司，秋山幸司，川崎正博
（富士通）

1．はじめに

知的情報検索システムIRISの大きな目標の一つは，テキスト文の内容理解に基づくテキスト情報の検索である［1］。そのため，IRISでは，テキスト文の内容を表現する言語を定め，その言語で表現されている内容同士の意味の類似性を判定する手続を導入している。本稿では，この言語及び類似度判定手続の概要を説明し，この判定手続がどれ程人間の知的判断と同じか，異なるかについて報告する。

2．内容照合過程の分析

内容の類似度判定の手続の実現方法を考えるため，人間の内容照合過程を分析した。一人の被験者に次のようなことをしてもらった。アンケート調査などにより収集した質問文約 360例のうち，典型的だと考えられる質問文を18個選択し，この質問文に答えるようなテキスト文を約840 件のテキストベースから検索してもらった。次に，このテキスト文それぞれについて，それが本当に質問文に答えるようなテキストであるかどうかを判断してもらった。その結果，その判定は，白黒はっきりしたものではなく，何か慎然としたものであり，例えば，図１に示されるような結果になった。

質問文例：	最近発売された１６ビットパソコンはどのようなものか？	
内容一致テキスト	：	「日電が３２，１６ビット機　イメコンＭＳシリーズ　二五六Ｋ・ＤＲＡＭ採用実装効率など向上」
まあ一致	：	「富士通　ＦＭフェスティバル　７０社参加し盛大に開催」
かなり違う：		「明治生命　カシオと共同開発　携帯パソコンを大量導入」

図１　人間による内容判定例

図１の例では，質問文中の「16ビットパソコンが発売される」という命題が中心となり，内容の一致が判定されている。また，製品に関する記述がより詳細に一致する方が全体の一致度も上がると判定されている。即ち，単に「パソコン」というより「16ビットのパソコン」といった方がよりよくマッチする。さらに，具体的製品名「FM」がパソコンであるということが用いられて一致度が判定されている。「発売される」に関しても「フェスティバル」が「発表」や「発売」の場所になり得るという判断が働いているようである。

以上のような分析を各例について行った結果，質問文に含まれる「何が何をどうした」という情報が如何に一致しているかによって，テキストの内容一致度が判定されていると判断できることが判った。また，その際には，内容的に重要となる要素，組織体や製品など，に焦点があり，他の要素は無視される傾向にあることも判った。そこで，IRISの内容一致度判定モジュールである内容マッチャーのプロトタイプでは，入力意味構造のうち，内容的に重要となる意味要素に関してのみ内容マッチングを行い，内容マッチングの結果は，各意味構造の要素の一致に依存するような数値で表わすようにした。

3．内容マッチャー

内容マッチャーの入力である質問文の意味構造とテキスト文の意味構造は，ともに第２節で分析した内容マッチング上重要となる要素以外の情報も担っている。そこで，内容マッチャーでは，これらマッチング上重要となる要素だけに入力情報を絞り込むため，文の内容表現言語を定め，内容マッチャーの最初のフェーズで入力意味構造をこの言語に合うように標準化する。その後，内容マッチャーは，この標準化された意味構造同士の一致度を計算し，最後にある定数（閾値）を超えない一致度しか持たないテキストは，内容的にまったく一致しないものと見なし，検索集合から取り除いている。

3．1　内容表現言語

［1］で示されるように質問文やテキストの意味構造は内容モデルによって規定されている。このモデルは，格フレームに準じた述語の引数パターンや名詞的実体間の依存関係をモデル化したものである。これらの引数パターンや依存関係は，内容表現言語では，内容マッチング上重要なものだけに絞られる。引数パターンは，主に，主格や目的格に相当するような引数だけに限定される。依存関係は，現在のところ，そのまま内容表現の方に移される。

内容表現言語は，大別して，述語論理表現の部分と実体表現の部分に分けられる。述語論理表現の部分は，第２節で示した「何が何をどうした」という命題に相当するものを論理積結合した構造で表わされる。実体表現の部分は，この命題記述中の「何」の部分に相当し，それが有るクラスを表現する内包的表現（例えば，「16ビットのパソコン」）であるか，１つのインスタンスを示す外延的表現（例えば，「FM16β」）であるかの区別を持っている。

述語論理表現の述語は，主に，内容モデル中の述語オブジェクト［1］に相当し，その引数は，前述したように主な格関係に限定されている。実体表現は，主に，内容モデル中の名詞的オブジェクト［1］に対応し，その名詞的オブジェクトが持つ依存関係は，そのまま持っている。このように，内容モデルの述語オブジェクトは，内容表現言語の述語に，名詞的オブジェクトは，内容表現言語の実体表現にほぼ対応するが，中には，内容モデル上は名詞的オブジェクトであり，内容表現上は述語になるものもある。この代表的なものとしては，"属性"に属するオブジェクトがあり，これらは，内容表現上は述語として扱われる。

Content Retrieval Aspect of IRIS
by Kenji Sugiyama, Kohji Akiyama
and Masahiro Kawasaki,
Fujitsu Laboratories Ltd. and Fujitsu Ltd.

3．2　内容マッチング処理

上で述べたように，内容表現言語が大きく述語論理的記述部と実体記述部の2つに分れているので，内容マッチング処理もそれぞれの記述に対応して2つに分けて実現されている。述語記述部に関する内容マッチング処理では，一階述語論理による前向き推論による一致度の計算を行なう。

図2に示されるように，「ある組織体Xがある製品を発売する」ということは「ある組織体Xがその製品を作成する」という命題が成立している可能性があるという世界知識［1］を使って前向き推論が進められる。各推論のステップでは，その命題間関係がどれ程密接に関連するかをモデル化したcertainty factorを持っており，各ステップでのcertainty factorをかけ合せることにより，初期命題と結論との隔たりを判断するようになっている。

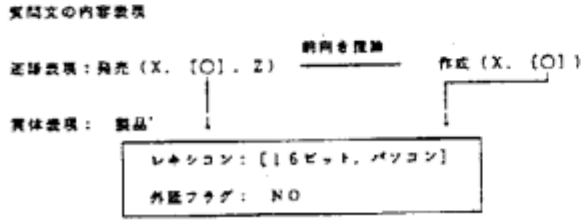

図2　推論

述語記述部の内容一致度は，結論命題とテキスト内容命題の一致度と上の隔たり値との積になっている。命題同士の一致の条件は，述語名が同じで，対応する引数が内容的に一致することであり，一致度は，各引数の内容一致度の和で表わされる。各引数は，prologのunificationとほぼ同様に一方が変数なら，unifyすることによって内容一致すると判定され，定数同士，すなわち，両方とも実体表現になっているのなら，それらが内容一致するかどうかによって引数の内容一致性が決定される。

実体表現の一致は，外延的表現同士であれば，レキシコン自体が一致すれば内容一致すると判定され，一方が内包表現で他方が外延的表現であれば世界知識の中の外延的知識［1］が使われ内容が判定される。内包表現同士の場合はレキシコン自体が一致するか，あるいは，世界知識の中のシソーラス的知識［1］によって一致性が判定される。内容一致度は，レキシコン自体同士の一致のように直接的に一致する場合は大きな値，知識を利用して一致性を判定するような間接的な場合は知識利用のステップが長い程小さな値になるようにしている。

4．評価

現在，内容マッチャーのプロトタイプに組み込まれた世界知識の量は，命題間知識が25ヶで，その他の実体に関する知識は，約220ヶ程度である。本プロトタイプの能力を評価するため以下のような実験を行なった。実験で使用した質問文例は2節で述べた18例である。テキストベースの方は，約840件のテキストベースから第1ステップとして抽出されたテキスト文72件である。それぞれの文例は，質問文解析モジュール及びテキスト文解析モジュールによって解析され，質問文例の方はすべて正しい意味構造に変換されている。テキスト文例の方は約85%は，完全あるいは不完全ではあるが，正しい意味構造に変換されているが，残り約15%は，誤った意味構造が作り出されている［2］。

次に，各質問文に対してどのテキストが検索されるべきかを検討した。これは，ユーザの意図や人によって異なる可能性があるので3人の被験者にそれぞれ独立にどのテキストを選ぶべきかを検討してもらった。その結果，3人とも一致する場合もあれば，一致しない場合もあり，かなり随意性があることがわかった。そこで今回の評価では，多数決により，正解とすべきテキスト集合を決定した。

情報検索システムの能力の評価としては，適合率，再現率がよく用いられるが，内容マッチャーはテキスト集合を見つけるというより，内容一致度の計算によりテキストの関連性の順番決定を行うので，順序集合に関する評価方法(1)再現率－適合率曲線(2)正規化再現率，適合率による評価［3］を行った。(1)による典型的質問文に関する評価グラフを図3に示す。グラフがより右上にある方がシステムの検索能力が高いことを示している。また，(2)による平均的な正規化再現率と適合率は，それぞれ0.85と0.82である。

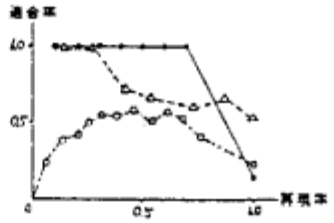

図3　再現率－適合率曲線
（1つの曲線が1つの質問文における再現－適合率の相関を表す）

5．おわりに

本稿では，内容マッチングの概念とその能力について報告した。今後は，さらに多量のデータによる評価が必要となる。その際，情報検索評価の際に入り込む人間的随意性を取り除く適切な手段が必要となる。

また，本稿で述べた能力の評価以外にも，検索スピードやコストの面からの検討も必要となる。その際には，内容マッチャー単独ではなく，キーワード検索式自動生成エキスパートを始めとするIRIS全体としての評価が要求される。

謝辞　本研究は，ICOT研究の一環として行なわれた。ここに記して感謝する。

参考文献

［1］杉山他：“自然言語に基づく情報検索システムIRIS”，情処NL研58-8，pp.1-8，1986

［2］伊吹他：“自然言語インターフェイスとしてのIRIS”，情処34回全国大会4X-8，1987

［3］伊藤：“情報検索”，昭晃堂，ソフトウェア講座19，p.174，1986