検索対象テキストＤＢ自動決定法の検討

1C－10

田中智博　松尾比呂志　木本晴夫

ＮＴＴ情報通信網研究所

１．はじめに

近年、様々な分野の文書が電子化され、様々なＤＢが利用できる環境になりつつあるが、どのＤＢを検索するかは、利用者が各々のＤＢの概要、あるいは、分野名を参考に決定しているのが現状である。しかし、対象とするＤＢの種類が多くなれば、ＤＢ選択に多大の労力が必要となる。そこで、我々は、各々のＤＢの内容を圧縮した情報を、ＤＢ決定用知識としてもたせ、これを用いて、検索要求文（「自動車を大量に輸出した年は？」などのような１文）から自動的にＤＢを決定する方法を検討した。

２．基本的な考え方

・ＤＢ決定法

各テキストＤＢと、その内容を圧縮した情報（ＤＢ内容と呼ぶ）との対応関係を記述したＤＢ決定用知識をあらかじめ作成しておき、検索要求文から抽出された情報と、ＤＢ内容との照合を行うことにより、テキストＤＢを決定する。

・ＤＢ決定用知識

新規ＤＢの追加にも容易に対処できるように、各テキストＤＢ内の文書群から、自動的にＤＢ内容を抽出して作成する。

・照合用情報

文書および検索要求文から抽出する情報としては、自立語（名詞、用言）、ＫＷとなる単語（従来より用いられているＫＷの自動抽出手法（例えば[1]など）による）の２つの場合について、ＤＢ決定用知識側、検索要求文側のそれぞれに適用し評価を行う。

・シソーラスの利用

シソーラスは、検索や分類において有効であるとの報告（例えば、[2],[3]など）がなされており、ここでもシソーラスを利用する。具体的には、文献[2]で用いられている意味属性体系（一種のシソーラス）を使用し、単語を意味属性に変換して使用することとする。この体系は、一般名詞に対して約2,800、固有名詞に対して約130の意味属性を設定している。なお、１単語に対して複数の意味が考えられる場合には、考えられる複数の意味属性を付与し、用言に対しても用言性名詞と同様な意味属性を付与してある。

３．ＤＢ決定手法

上記の考え方を基に検討を進めているテキストＤＢ自動決定の処理構成を図１に示す。

３．１　ＤＢ決定用知識作成

各テキストＤＢごとに、文書群中の文書から、単語を抽出する。単語の抽出方法としては、文献[1]に述べられているＫＷ抽出処理（並立表現語、連体修飾語、不要語などの処理）と、単に形態素解析により、自立語のみを抽出する方法の２つを用意する。抽出された単語に対して意味属性を付与し、テキストＤＢごとに各意味属性の出現頻度をカウントする。なお、単語に複数の意味属性が付与される場合には、それら全てをカウントする。そして、全てのテキストＤＢに対して、意味属性の出現頻度がカウントされた時点で、各意味属性ごとに、各テキストＤＢに対する出現頻度を得点に変換し、ＤＢ決定用知識とする。すなわち、ここでのＤＢ決定用知識は、各テキストＤＢに対する得点が記述された意味属性の集合となる。ＤＢ決定用知識の一例を図１に示す。

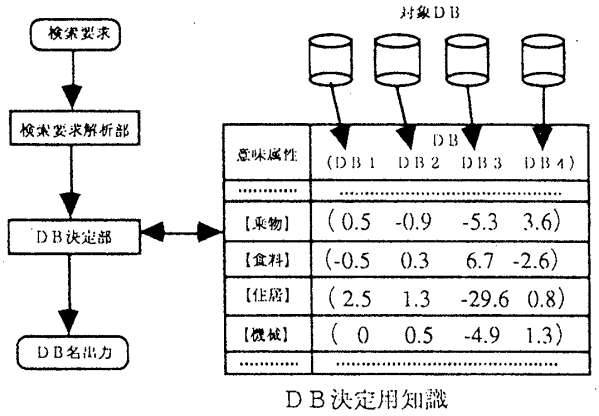

図１　ＤＢ自動決定処理構成

頻度から得点への変換式としては、文献[2]で示されている次式を用いた。

$$Yjk = (Fjk - Mjk) * |Fjk - Mjk| / Mjk$$

$$Mjk = (\sum_{j=1}^{n} Fjk / \sum_{j=1}^{n}\sum_{k=1}^{m} Fjk) * \sum_{k=1}^{m} Fjk$$

ここで、Yjkは、意味属性jのテキストＤＢkにおける得点、Fjkは、意味属性jのテキストＤＢkにおける頻度、Mjkは、意味属性jがＤＢに依存せずにランダムに出現した場合のテキストＤＢkにおける頻度（理論度数）、mはテキストＤＢ数、nは意味属性の異なり数（約2,930）を表す。

この式により、得点が、Fjk - Mjk＞0（ランダムに出現する以上の頻度で出現してる）場合には高く、Fjk - Mjk≦0（ランダムに出現する以上の頻度で出現していない）場合には低くなるように計算される。

例えば、図１の意味属性【乗物】の場合、ＤＢ４の得点が他のＤＢの得点に比べて高くなっており、検索要求文中にこの意味属性が出現した場合、ＤＢをＤＢ４と特定すべき手掛かりとなることを示している。

３．２　検索要求文解析

ＤＢ決定用知識作成同様、ＫＷ抽出処理、自立語抽出処理の２つを行い、抽出された単語に対して意味属性を

An Automatic Database Selection Method for Information Retrieval

Tomohiro TANAKA,Hiroshi MATSUO,
Haruo KIMOTO
NTT Network Information Systems Laboratories

付与する。

3．3　ＤＢ決定

検索要求文から得られた意味属性に対して、ＤＢ決定用知識から各ＤＢに対する得点を付与し、検索要求文全体における全ての意味属性の得点を各テキストＤＢごとに合計する。１単語に対して複数の意味属性が付与されている場合は、その単語に付与された意味属性の得点の平均を求め、その単語の意味属性の得点とする。

検索要求文全体について得られた各テキストＤＢの正の値を持つ得点中もっとも大きいもので、各テキストＤＢの正の値を持つ得点を割り、その検索要求文に対する各テキストＤＢの正の値を持つ得点が０～１の範囲で分布するようにする。なお、負の得点を持つＤＢは、除外する。

結果出力においては、０～１の間で足切り値を設定し、その値以上の得点をもつ全てのテキストＤＢを、適合するテキストＤＢとして出力する。

例えば、図１に示すＤＢ決定用知識を用いて計算すると、検索要求文から得られた意味属性が【住居】【乗物】【機械】（【乗物】と【機械】は１単語から得られた）であった場合、１単語から得られた【乗物】と【機械】の得点を平均化した得点（0.3 -0.2 -5.1 2.5）と、【住居】から得られた得点を合計して得点（2.8 1.1 -34.7 3.3）が得られ、各ＤＢの正の値を持つ得点中最も大きな値で、各ＤＢの値を割って、（0.85 0.29 --- 1.00）が得られる。ここで足切り値として0.8が設定されていれば、適合するＤＢとしてＤＢ４、ＤＢ１を出力する。

4．評価・考察

4．1　評価実験

新聞の分野をそれぞれ独立したテキストＤＢと考え、新聞記事（１９９２年分）の７分野（科学、経済、社会、スポーツ、政治、文化、事件）からランダムに抽出した各２００記事（１記事約２００文程度）を対象テキストＤＢとし、約４０名の男女（２０～４０歳）により作成した検索要求文（６６文）を用いて実験を行った。検索要求文に対して、作成した人及び、作成した人以外の３名により適合すると思われるＤＢを決定してもらい（複数ＤＢの決定を許す）、２名以上が決定したＤＢを正解ＤＢとした。表２に示す条件に対して、足切り値を0.5～1.0まで0.05刻みで設定し、以下に示す再現率と適合率を求めた。

表１　評価実験条件

		ＤＢ決定用知識作成 KW抽出	ＤＢ決定用知識作成 自立語抽出
検索要求解析	抽出 KW	条件１	条件２
	抽出 自立語	条件３	条件４

再現率＝(RETrel／rel)*100、適合率＝(RETrel／REL)*100
RETrel＝システムが出力した正解のＤＢ数、
rel＝正解のＤＢ数、REL＝システムが出力したＤＢ数。
実験結果を図２に示す。

4．2　考察

条件３（ＤＢ決定用知識側がＫＷ抽出、検索要求文解析側が自立語抽出）のとき、精度が良くなっている。これは、文書群を対象としたＫＷ抽出の場合、語の絞り込みにより、不要な語が削除され、比較的文書群に特異な語を抽出できているが、文を対象としたＫＷ抽出の場合には、元の語数が少ないために、必要な情報まで削除しているためである。

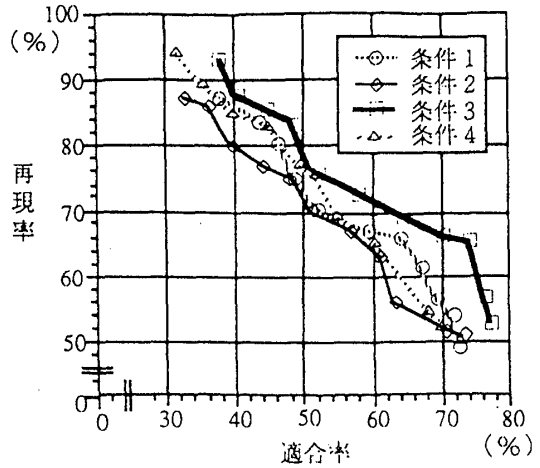

図２　評価実験結果

条件３において、失敗している主な原因とその対策について以下に述べる。

・検索要求文から、自立語のみを抽出しているために、不要な（ＤＢ決定において負の作用を及ぼす）単語を抽出していることに起因する失敗。例えば、検索要求文「人工衛星の軍事以外の利用を知りたい。」(正解ＤＢとしては、科学となるが、単語「軍事：意味属性【軍事】」の影響により、「政治」を出力している。)

検索要求文において単に自立語全てを抽出するのではなく、単語間の論理関係（上記の例では、NOTの関係）等を考慮して、抽出する単語を選定する必要がある。

・１単語に複数の意味属性が付与されている場合、得点計算において、単に平均をとっていることに起因する失敗。例えば検索要求文「自動車の輸出について知りたい。」（正解ＤＢ：経済）の場合、「自動車」の意味属性として【車】(得点の高いＤＢは経済、事件)、【スポーツ】（得点の高いＤＢはスポーツ）の２つが付与されており、平均をとると得点の高いＤＢがスポーツとなる。

意味属性を付与する際に、構文構造等の情報に基づいて、付与する意味属性の絞り込みを行う必要がある。

5．おわりに

本報告では、テキストＤＢ自動決定手法の検討として、ＤＢごとにＤＢ決定用知識を作成し、意味属性の頻度情報に基づく手法の評価を試みた。その結果、ＤＢ決定用知識作成側では、ＤＢ中の文書群からＫＷを抽出し、検索要求文側では、自立語全てを抽出する方法が、有効であることが分かった。今後は、検索要求文から、ＤＢ決定に有効な単語の抽出を行う方法、意味属性を付与する際の絞り込みの方法を検討するとともに、ＤＢ側の汎用性を考慮しながら、頻度情報以外の情報によりＤＢを決定する手法についても検討を行う。

参考文献

[1] 木本：「日本語新聞記事からのキーワード自動抽出と重要度評価」、信学会論文誌Vol.J74-D-I No.8 (1991)
[2] 河合：「意味属性の学習結果にもとづく文書自動分類方式」、情処学会論文誌Vol.33 No.9 (1992)
[3] 松尾、内野：「意味属性に基づくテキストベース検索方式」、情処学会論文誌Vol.32 No.9 (1991)