E-8

知的情報検索のための概念学習方式

The Method of Concept Learning for Intelligent Information Retrieval

伊藤 俊介† ITOH Shunsuke　藤井 啓彰† FUJII Hiroaki　渡部 広一† WATABE Hirokazu　河岡 司† KAWAOKA Tsukasa

1．はじめに

急速な WWW の流行に伴い，氾濫する情報の中から，必要な情報を取得するために「情報検索」という研究が盛んに行われ，実用化されている．しかし，現状のサービスは，検索システムをよく使用する，慣れたユーザでないと目的の情報をすぐには得られない．また得た情報をすぐに理解することは，ユーザにとってかなりの負担となっている．この情報検索の問題点を語の知識を蓄えた概念ベース[1]を用いて改善し検索システムの知的化を行うことが，本研究の課題である．

2．目標とするシステム

理想の情報検索とは，検索者の検索要求を的確に満たすと同時に，検索者にあらたなる発見や想起を促す関連情報を提供することであると考える．本研究では，図1のように概念ベース，シソーラス[3]に存在しない語（未定義語）の検索要求に対して，その語の語彙体系（その語が属すると思われるシソーラスのノード）を提示し，検索者の理解を支援することができるシステムの開発である．

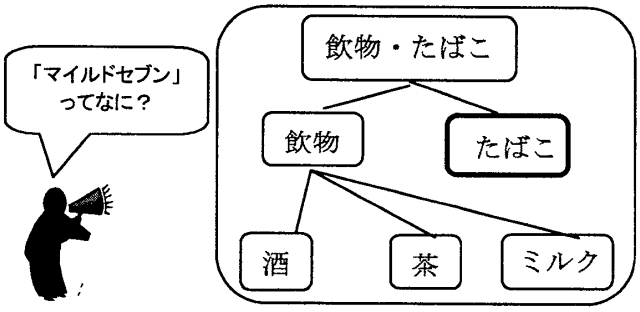

図1　理解支援システム

2.1 概念ベース

概念ベースとは，語（概念）に複数の関係のあるような語（属性）を持たせた知識ベースであり，約10万語が格納されている．

概念は，属性と重みの対の集合として表されている．ある概念を A，その i 番目の一次属性を a_i，重みを w_i とする．

$$A=\{(a_1,\ w_1),\ (a_2,\ w_2),\ \cdots,\ (a_n,\ w_n)\}$$

使用した概念ベースはこの重みを五段階（概念 A と関係が高いと思われるものから順に 10,9,4,3,1）にふったものである．

2.2 未定義語の概念化

検索対象文書群において，未定義語の検索要求語一語で表記一致検索を行い，関連文書を取得する．

取得した関連文書群に含まれる全ての単語について $f\cdot idf$ 値[2]を加算し，高得点の語を未定義語の属性として採用する．

$$w_t^d = tf\ (t,d)\cdot idf\ (t)$$

$$idf\ (t) = \log \frac{N}{df\ (t)} + 1$$

$tf\ (t,d)$項は，ある文書中 d に出現する索引語 t の頻度である．$df\ (t)$項は索引語 t が出現する文書数である．また N は文書空間の全文書数である．

$tf\cdot idf$ とは，単語の頻度と網羅性に基づいた語の重み付けである．tf は頻度情報を用いて適切な属性を集めたものであり，idf は特定性情報を用いて特徴付けた属性を集めたものである．表1に未定義語を概念化した例を示す．

表1マイルドセブン(未定義語)の属性

属性	JT	ライト	たばこ	銘柄	F1	…
重み	97	86	27	19	17	…

2.3 シソーラスのノードの概念化

概念化により，未定義語を概念として扱えるようになったが，比較対象であるシソーラスのノードはそのままの状態では，概念化が行われていないため比較が困難である．そのためシソーラスのノードも概念化を行う．

図2のようにシソーラスの各ノード（2710 個）に対してノードに属するリーフを全て概念ベースで参照し，属性とその重みを足しあわせて属性集合を取得する．この属性集合にノード名を概念ベースで参照して取得した属性を加えたものをノード属性とする．ただしノード名が概念ベースに存在しない場合はリーフによる属性取得のみを行う．

すべてのノードにおいてこの処理を行いシソーラスの「tf 重み付きのノードの属性」を構築する．次にこの構築したシソーラスの全属性内で，idf 重み付けを行い，重みに特定性情報を付加する．これを「$tf\cdot idf$ 重み付きノード属性」とする．

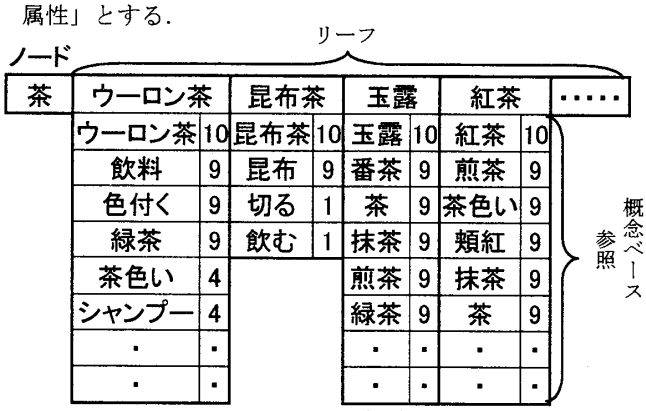

図2 ノードの概念化方法

2.4 未定義語とノードとの関係推定

未定義語の属するノードを特定する手法として，構築した未定義語の属性と，概念ベースにより概念化したシソーラスの各ノード属性と関係推定を行う．関係推定の方法として重み付き関連度計算方式($ChainW$)[1]を使う．未定義語の概念を $A=\{(a_i)|i=1\sim L\}$，シソーラスのノード属性を $B=\{(b_j)|j=1\sim M\}$ とし，その重み付き概念連鎖関連度 $ChainW(A,B)$を求める方法を示す．一次属性 a^1_i と b^1_j の重み付き一致度 $MatchW(a^1_i,\ b^1_j)$を次のように定義する．u_i，v_j は a_i，

† 同志社大学大学院 工学研究科
Graduate School of Engineering, Doshisha Univ.

b_jの重みをL，Mは概念A，Bの属性数を表す．

$$MatchW(A,B) = (s_A / n_A + s_B / n_B)$$

$$s_A = \sum_{a_i=b_j} u_i \qquad s_B = \sum_{a_j=b_j} v_j$$

$$n_A = \sum_{i=1}^{L} u_i \qquad n_B = \sum_{j=1}^{M} v_j$$

I　属性の少ない方の概念をAとし（$L \leqq M$），概念Aの属性の並びを固定する．

II　概念Bの各属性を対応する概念Aの各属性との重み付き一致度$MatchW(a_i, b_j)$の合計が最大になるように並び替える．ただし，対応にあふれた概念Bの属性$(b_i | j=L+1,\cdots,M)$は対応をとらないものとする．

III　重み付き一致度から$ChainW(A,B)$を求める．

$$ChainW(A,B) = (s_A / n_A + s_B / n_B)$$

$$s_A = \sum_{i=1}^{L} u_i MatchW(a_i, b_{x_i}) \qquad n_A = \sum_{i=1}^{L} u_i$$

$$s_B = \sum_{i=1}^{L} v_i MatchW(a_i, b_{x_i}) \qquad n_B = \sum_{j=1}^{M} v_j$$

上記のような重み付き関連度計算を，未定義語の属性と全てのノード属性との間で行い，もっとも値の高かったノードを，対象の未定義語が属すべきシソーラスのノードとする．概念化した未定義語およびシソーラスノードの属性に存在する未定義語は，計算から除外している．

3. 検索対象および評価方法

検索対象は毎日新聞 94 年度版 CDROM より一年分の新聞記事を対象とした．評価については，人手で約 400 個のテストセットを作成した（表２）．テストセットは未定義語のみで構成されている．

表２ テストセットの一部

バッハ	ソニー	阪神大震災	パジェロ
C型肝炎	バファリン	青函トンネル	……

テストセットの未定義語での検索要求に対して，どれだけの割合で正しいシソーラスのノードを導くことができるかを評価する．未定義語が属すべきノードを判定するときに，正答とすべきノードが複数あり，一意に属するノードが確定しないことが多い．よって，結果が正答か否かは人手で評価した．

もう一つの評価法として，シソーラスの各リーフの表記（2710 個）を未定義語に見立て，その語の属性を概念ベースより取得する．概念ベースの属性は精錬作業が行われており属性に雑音が少ない，つまり理想的な属性である．これを未定義語の属性としてシソーラスのノード特定を行い，どれだけ元々所属したノードを返答することができるかを評価した．この評価法は「未定義語とノードとの関係推定方式」の部分のみを評価を目的としている．

4. 評価

まずシソーラスのリーフを利用した評価法の結果を表３に示す．ノード属性および未定義語に見立て，その語の属性を概念ベースより取得したリーフ属性ともに属性数にバラツキがあるが，構築したすべての属性を使用している．

属性が十分良い（雑音が少ない）状態では重み付け方式を駆使しなくても，よい結果が得られた．

表３ リーフによる自動評価の正解率

	ノード属性tf	ノード属性$tf \cdot idf$
リーフ属性tf	61.444%	61.655%
リーフ属性$tf \cdot idf$	61.411%	61.601%

次にテストセットを使用し，シソーラスの各ノードと関係推定を行い，結果を人手で評価したときのデータ示す．

未定義語の属性を構築するときには，多くの雑音が属性に混入する．そのため構築したすべての未定義語の属性を使用すると，雑音が信頼性のある属性を駆逐してしまい，精度が 0%に近い値となってしまう．そのため未定義語の属性に関しては，重み上位 20 個打ち切りで調査を行った結果を図３に示す．

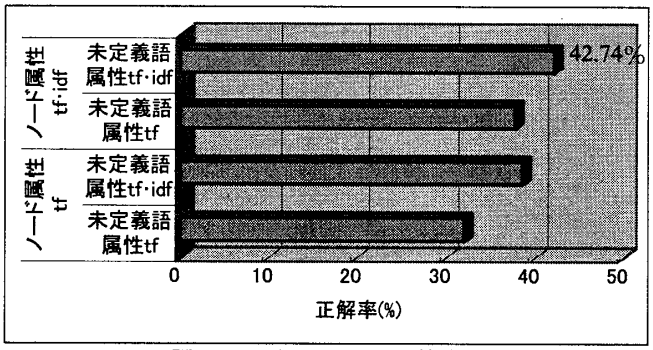

図３ テストセットの正答率

属性に雑音が多い状態では重み付け方式を駆使することによって，正答率が向上した．

5. 考察

未定義語の属性に雑音が多い状態では$tf \cdot idf$重み付けを利用することが正答率の向上に有用である．

未定義語の属性の精度が十分良い，つまり雑音が少ないときには，重み付けの方式などを色々と駆使しなくても，良い結果がでることが分かった．

したがって重み付け方式を改良していく前に，未定義語の属性の精度向上を行っていくことが，今後のシステムの能力向上に有効であることが分かった．

さらなるシステムの精度向上には他の手段(ニューラルネットワーク，決定木学習，遺伝的アルゴリズム)などの技術を組み合わせることが有効だと考えられる．今後更なる精度向上を行うため，これらの方法について模索していきたい．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

[1]渡部広一，河岡司：常識的判断のための概念間の関連度評価モデル，自然言語処理，Vol.8，No.2, pp.39-54, 2001.

[2]徳永健伸:情報検索と言語処理，東京大学出版会, 1999.

[3]NTT コミュニケーション研究所，日本語語彙体系，岩波書，1997