6E−7

文書の整理に関する一考察*

福永 博信†
NTTヒューマンインタフェース研究所‡
fuku@ntthli.ntt.jp§

1 はじめに

近年の新しいメディアや計算機ネットワークの普及とともに計算機可読文書の流通量は大きく増加している。しかし現状では、蓄積したものを検索して読むという低次元の利用しかされていない。これは、(1)利用者の要求の多様さ(2)言語処理技術の未成熟(3)処理用辞書などの未整備などの原因が考えられる。本稿では、文書の高度利用に関して整理という観点から考察する。

2 文書の利用

文書は、人間相互間の情報伝達の手段として用いられる。文書のライフサイクルを次のようにとらえることができる。発信者の知的活動の結果として作成され、発信される。受信者が受信した後、利用,蓄積あるいは廃棄する。蓄積された文書は、必要に応じて知的活動に利用され、場合によっては廃棄あるいは変形して再蓄積される。このモデルに従えば、文書をそれらの状態間で移動させる作業は次のように考えることができる。

- 受信した文書を、必要な文書がどうかを判断して廃棄するか蓄積するかを決定する作業が分別であり、人間の知的活動の一部ともとらえられる。これを自動的に行なうのがfiltering技術である。
- 蓄積した文書の中から必要とするものを探す作業が検索であり、これも人間の知的活動の一部ととらえることもできる。これを精度・効率良く実行するための技術が文書検索技術である。

人間の純粋な知的活動である利用の部分を次のように考える。人間は文書を読んで解析し、自分の世界モデルと照合しながら解釈し、その結果を自分の世界モデルの中に知識として格納あるいは既存の知識を更新する。すなわち、人間は自然言語から自分の世界モデルの中に投射して利用している。

filteringや文書検索の技術は、文書をどの状態(廃棄,蓄積,利用)にするかという部分に関しては人間の知的活動を支援しているが、形態は自然言語で記述された文書のままである。従って、人間が利用する際には上述の作業を行なうことになる。本稿では利用時の解析,解釈,照合部分を支援する技術の研究を提案する。

3 概念・知識・世界モデル

人間は自分の経験を通して獲得した知識から構成される世界モデルを持っている。知識を概念相互間の関係ととらえ、世界モデルを概念の節とする網である意味ネットワークでモデル化すると、知識獲得は次のように定義できる。

定義1：知識獲得
知識の獲得は、自分の世界モデルの網を拡張・修正することである。■

一つの知識は、数個の概念とそれらの関係であり、その知識を世界モデルに加える際の概念の新規導入・削除,概念間関係の追加・変更・削除が知識獲得であると定義する。獲得された知識は単一の知識として網に組み込まれ、その量の増加にともなって網は大きくなる。このモデルにおいては、知識相互間の関係は別の知識によってのみ言及されるため、類似事項の推論・類推は全て処理系に負わせることになる。一方人間は、情報処理に際して無秩序に発達した網の状態の世界モデルを利用しているとは考え難い。そこで、知識整理について次のように定義し、人間の知識整理を仮定する。

定義2：抽象知識
より上位の概念相互間の関係で記述した知識をその知識の抽象知識と定義する。■

定義3：知識整理
知識を整理することは、抽象知識に従って知識を整頓することである。■

仮定1
人間は知識を獲得した後、整理した状態で利用する。■

すなわち、人間は知識を束ねた状態で管理しているという仮定である。以下、この仮定に基づき、この知識整理の方法について考察する。

4 知識獲得・整理

3節に従うと、知識は数個の概念とそれら相互間の関係であるから、文書から知識を獲得する作業は、言語処理技術を適用して

*A Study of Text Ordering
†FUKUNAGA,Hironobu
‡NTT Human Interface labs.
§fuku%ntthli.ntt.jp@relay.cs.net

(1) 文章から概念を抽出する

(2) 文章から概念間の関係を抽出する

の2つを行なうことと規定される。従って、この種の知識獲得器の能力は、

1.1 取り扱える概念の種類

1.2 文書中の表現と概念との照合精度

2.1 取り扱える概念間関係の種類

2.2 文書中の表現と概念間関係との照合精度

の4つで測ることになる。

知識整理は、獲得した知識を抽象知識で束ねることである。その作業は、

(3) 上位概念・上位関係との照合

(4) 抽象知識の枠で整形する

の2つを行なうことと規定される。従って、この種の知識整理器の能力は、

3.1 概念，概念間関係体系の充実度

3.2 上位の概念，概念間関係との照合精度

4 抽象知識と人間の知識整理モデルとの間の距離

の3つで測ることになる。

上述の知識獲得・整理器の動作手順は次のようになる。

1. 言語処理
   文書に対して言語処理を施し、
   知識 = (概念 + 概念間の関係)
   を抽出する。
2. シソーラス処理
   シソーラス辞書を参照して抽出した概念の上位概念を得る。
3. 格納処理
   抽象知識に従って、知識を枠に適合させ格納する。

1.1, *1.2*, *2.1*, *2.2*は言語処理系、*3.1*は概念体系、*3.2*は概念体系と抽象知識、*4*は抽象知識の完成度に依存する。

5 抽象知識獲得

抽象知識は知識整理に必須であるが、それを準備することは比較的困難である。それは、次のようなことに起因する。

(1) 知識獲得系と同一の概念体系，概念間関係で記述する必要がある。

(2) 多層をなす概念の選択が困難。

そこで4節と同一の言語処理系とシソーラス辞書を用いて、次のような手順で構築する方法を提案する。

1. 文書に対して言語処理を施し、概念と関係で表される知識に分解する。
2. 知識の集合に対して、ある概念を中心に整頓する。
3. シソーラス辞書を参照して、整頓して同一関係にある概念群に適合する上位概念があれば、抽象知識として獲得する。
4. 抽象知識と文書から獲得した知識と同様に扱いながら、2,3を繰り返す。

シソーラス辞書とは、「言葉」の体系を収録した辞書であり、厳密には「概念」の体系とは異なるが、近似的に「概念」の体系として用いる。従って、同義語などの取り扱いなどに注意を要する。

6 言語処理系と世界モデル

2節の文書利用のモデルにおける解析,解釈,照合の作業は利用者の世界モデルを参照して行なわれる。4節の知識獲得・整理器において、世界モデルの参照は言語処理系が行なう。従って、計算機中に利用者の世界モデルに相当するものを用意しなくてはならない。その際、個々の知識全てを用意・利用するのは現実的ではない。シソーラス辞書と抽象知識をその代替として利用する方が現実的である。

7 おわりに

計算機で取り扱える文書データは増大し、比較的容易に入手できるようになっている。しかし、文書を利用した人間の純粋な知的活動の前に行なわなければならない仕事の量は決して小さくない。人間の整理モデルと近い状態に整理された形態で文書を提示・保存すれば人間はより高速にあるいは高度な知的活動ができるという発想のもとに文書整理に関する研究を提唱した。また、知識獲得・整理器の能力を評価するためのポイントを示したが、今後は、その評価の尺度を考えるとともに、本モデルの有効性を確かめるための実験を行う予定である。

謝辞

有意義な議論を頂いた 中川透主幹研究員をはじめとするヒューマンインタフェース方式研究部の方々に感謝いたします。

参考文献

福永： 「文書の整理に関する一考察」；情処NL研資料 '91.5(発表予定)