ハイパーテキストデータベーススキーマ作成のためのノードの分類法

5Q－9

西村憲二　石川佳治　植村俊亮
奈良先端科学技術大学院大学 情報科学研究科

1　はじめに

ハイパーテキストシステムの短所であるデータ管理機能を補うために，データベースを融合して文書とその付加情報を管理するハイパーテキストデータベースの研究が行われている[2),3)]。

ハイパーテキストデータベースの設計方法は，大きく下向設計 (top-down design) と上向設計に分けられる[5)]。下向設計 (bottom-up design) は，まず最初にスキーマを設計し，それから航行モデルそしてインスタンスであるハイパーテキスト文書のノード／リンクを作成する。それに対して上向設計は，最初にハイパーテキスト文書を作成し，それをインスタンスとするようなスキーマを作成する。どちらの設計方法も必要とされている。

本研究では，上向設計に注目する。そして，ハイパーテキスト文書から実体を抽出するためのノードの分類方法について論じる。

図1に本論文で使用するハイパーテキストの例を示す。太い枠線のノードは索引ノードを表す。索引ノードはある基準に基づいて文書ノードを集めリンクを張ったノードである。

2　スキーマの上向設計

既存のハイパーテキストを活用するためにハイパーテキストから意味および構造情報を抽出してハイパーテキストデータベースを作成する上向設計が必要になる。スキーマの上向設計は以下のように行われる。

1. 実体の抽出
2. 航行設計（図2）
3. 概念モデル設計（図3）

なお，図2の記法は，RMD モデル[6)]を参考にしている。

手順２では１で抽出した実体と実際のリンク構造から航行図を作成する。手順３では，航行モデルを概念モデルに変換する。１の作業が終了すれば，２，３の作業はそれほど手間がかからないと考えられる。

3　ノードの分類

本研究ではリンク構造に基づいてノードを分類する。つまり，文脈的に類似したノードはリンク構造上でも類似した位置に配置されると考え，同じ実体に分類する（図1では，研究室→索引→メンバというリンク構造のメンバの位置にメンバに関するノードが配置される）。分類には，ＡＣＥ法[4)]を利用している。この手法について以下に説明する。

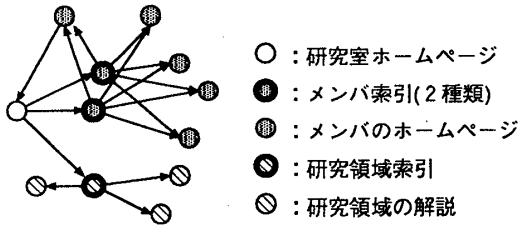

図 1: ハイパーテキストの例

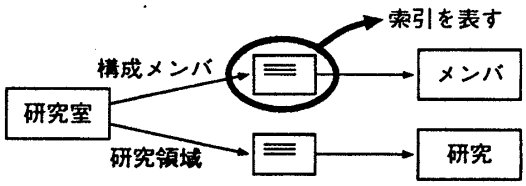

図 2: 航行モデルの例

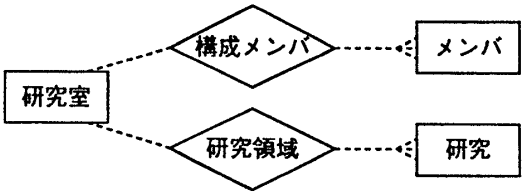

図 3: 概念モデルの例

3.1　ACE 法

ACE 法[4)]は，有向グラフとして表現できるハイパーテキストを対象とし，ハイパーテキストの高レベルな関連をとらえるためにもとのグラフ構造から集約と例外を生成するクラスタリング手法である。

グラフ $G=(V,E)$ の集約グラフ $G_A=(V_A,E_A)$ において，集合 $V_A=\{V_{A_1},\ldots,V_{A_n}\}$ は V の分割，集合 E_A は部分集合 V_A を接続するリンクである。ここでリンクが存在するかどうかは，元グラフのノード間のリンクの数による（すなわち，ある本数より多い場合に集約グラフのリンクが存在する）。また，ACE 法では例外を許しており，これを二種類に区別している。

inclusive link　元のグラフにリンクが存在するが，集約グラフ上の対応するノード間にリンクが存在し

Node Classfication for Scheme Construction of Hypertext Databases
Kenji NISHIMURA, Yoshiharu ISHIKAWA and Shunsuke UEMURA
Graduate School of Information Science, NAra Institute of Science and Technology (NAIST)

ない場合を指す。E_Iと記述する。
exclusive link 元のグラフにリンクが存在しないが，集約グラフ上の対応するノード間にリンクが存在する場合を指す。E_Eと記述する。

ACE法では，以下のコスト関数が最小になるようにクラスタリングを行う。

$$\text{Cost\_func}(G_A, G_I, G_E) = |V_A|+|E_A|+|E_I|+|E_E| \rightarrow \text{Min}$$

実際のハイパーテキストは，ノード数に比べリンク数が少なく情報を表すために必要なリンクがないなどリンクが構造情報を的確に表現していないことが考えられる。このことは，ACE法のみならずリンク構造を解析するすべての方法に共通の問題点であると思われる。そこであらかじめノードを分割し，その中でリンク構造に基づいた分類を行う。

分類精度の向上のために適用可能な二つの手法について以下で述べる。

3.2 索引ノードの選択

索引ノードは，ハイパーテキスト構造の中で航行機能上の意味をもち他の文書ノードから区別できる。そこで索引ノードを推測し，それらのノードは他のノード同じ分類に入れないようにする。また，索引ノード同士も区別されるため，それらが索引しているノードも区別されることが期待できる。

1) では，ノードの分類の前に索引ノードを選択している。その方法は，出ているリンクの数が平均値にしきい値を加えたものよりも大きいものを選択することである。また，テキストの大きさに対するリンクの数を指標にすることも考えられる。

3.3 ノードのグループわけ

あらかじめノードをいくつかのグループにわけ，同じグループに属するもの同士を分類するようにする。このグループわけの方法は，対象とするハイパーテキストの大まかな構造によって異なってくる。

例をあげると，

- 階層構造に注目する。対象のハイパーテキストに対して適当な階層構造を想定し，同じ階層に属するものを同じグループに入れる。
- 索引ノードに注目する。これらのノードが索引するノードを一つのグループとする。
- WWWであるならば，ＵＲＬを参照し，URLから階層構造を導く。

4 利用者による指定

図1に対して3章で述べた分類を実行した結果が図4 (a) であるとする。しかし，利用者がノードの模様に沿ったような分類をしたいと考えたとき，その意図を反映する必要がある。そのために利用者が分類を指定できるようにする。たとえば，図4(a) に対して2種類のメンバ索引を一つに分類すると指定した場合，分類結果は図4(b) になる。このように利用者は，自らの意図を反映しながら繰り返し分類処理を実行することで

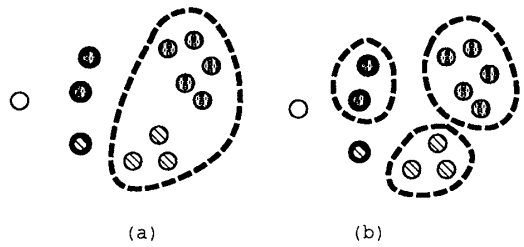

図 4: 分類結果

意図に基づいた分類を行える。利用者は分類結果をみて，ノードまたはノード群を移動させること，および以下のような指定を追加することができる。

分類指定 次の処理で同じ分類に入れるノード群を指定する。
リンク指定 上で指定したノード群の間のリンクを指定する。次の処理では，集約グラフにこのリンクが存在するとしてコスト計算する。
分類決定指定 ノード群がその状態で決定であることを指定する。次の処理ではそのノード群からなる分類に新しいノードを入れない。

他にも，同じグループに入れるノード群の指定，索引ノードの指定などが考えられるかもしれない。

5 まとめ

本論では，ハイパーテキストノードをリンク構造に基づいて分類する方法について論じた。今後は，実験結果を反映して，改良を進める。

参 考 文 献

[1] Rodrigo A. Botafago, et al., "Identifying Aggregates in Hypertext Structures", in *Proc. of the 3rd ACM Conference on Hypertext*, Dec. 1991, pp. 63–74.

[2] Chris Clifton, et al., "HyperFile: A Data and Query Model for Documents", in *VLDB Journal*, Vol.4, No.1, 1995, pp. 45–86.

[3] Yoshinori Hara, et al., "Hypermedia Navigation and Content-based Retrieval for Distributed Multimedia Databases", in *The 6th NEC Research Symposium Multimedia Computing*, 1995, pp. 1–15.

[4] Yoshinori Hara, et al., "Implementing Hypertext Database Relationships throuh Aggregations and Exceptions", in *Proc. of the 3rd ACM Conference on Hypertext*, Dec. 1991, pp. 75–90.

[5] 原 良憲 他, "ハイパーメディアプラットフォーム "雅 (みやび)" の概要", 情報処理学会研究報告, 92-DBS-90, Sept. 1992, pp. 29–38

[6] Tomas Isakowitz, et al., "RMM: A Methodology for Structured Hypermedia Design", in *Comm. of the ACM*, Vol.38, No.8, Aug. 1995, pp. 29–38