

分散環境での文学データベースの内容検索

1 X - 1

本行弘明 池口仁誠 三宅忠明 横田一正

e-mail: {hongyo, ikeguchi, miyake, yokota}@c.oka-pu.ac.jp

岡山県立大学 情報工学部

〒719-11 総社市窪木 111

1 はじめに

分散環境での複数の異種情報源を融合 / 統合するためには現在 QUIK メディエータシステム^[1, 5]を研究開発している。情報源の異種性は分散環境での自律性に基づくだけではなく、情報の抽象階層によっても生成される^[4]。本稿ではそのような情報源のひとつとして文学データベースを取り挙げ、QUIK メディエータシステムとしての適用可能性を議論する。

2 デアドラ伝説

ケルト(アイルランド)文学の国民的伝承物語としてデアドラ伝説(Deirdre Legend)がある。ゲール語には文字がなかったため、口承によって物語が伝えられ、数々の変種が存在する。1860 年以来発刊されたものが 51 種類、その他に口述筆記されたものが数多くある。この物語は時代と共に内容や語彙が変化しており、それは上の発刊や記録の時間的順序とは必ずしも対応していない。内容は、(1)誕生と予言、(2)養育時代、(3)出会いと駆け落ち、(4)帰国、(5)死、の 5 つから構成されており、数ページのものも数 100 ページのものも変わらない。

比較文学の立場からは、これら変種間の類似性や内容や語彙の変化を見つけ、変種の系統樹を作成することが大きな研究課題となっている。最近になって岡山県立大学で所蔵しているデアドラ伝説の資料の電子化を始め、われわれはそのデータベース化を現在検討している。デアドラ伝説のデータベース化はまだ始まったばかりであり、複数の研究者が分担してデータの入力や分析を行っており、ネットワークを介しての研究協力や利用が必要である。また口述筆記された物語は徐々に増加していくので、将来的には各地のデアドラ伝説データベースをネットワークで結び、比較文学研究を可能にすることを検討している。

Retrieval of Literature Databases in Distributed Databases
Hiroaki HONGYO, Noritaka IKEGUCHI, Tadaaki MIYAKE,
and Kazumasa YOKOTA
Faculty of Computer Science and System Engineering,
Okayama Prefectural University

3 文学データベースの構成

データベースは 1 つの物語に対して基本的に以下の異種の 3 種類から構成されておりそれぞれ特有の検索機能がある。

テキスト文書 キーワード検索

構造化文書 構造を意識したキーワード検索

ストーリ記述 内容検索

テキストに対しては通常の文字列検索である。構造化文書は構造は構文的な解析でほぼ十分で意味的なタグは不要である。したがってこれは、テキスト文書とほぼ同じと考えて良い。ただし部分構造をアクセスするために、タグ付きの巨大な複合オブジェクトとし、QUIK で表現する。

ストーリ記述については基本的に

- 完全なストーリ記述は不可能である
- ストーリには多様な解釈が存在する

という考えに基づいて構成することを考えている。したがってストーリは文献^[3]のようなボトムアップではなく、全体を 1 場面と考え、そこから 2 節の 5 つの場面に詳細化するようにトップダウンの記述を考えている。キーとなるのは「場面」の記述であり、以下のようにになっている。

場面 ID :

日時 :

場所 :

登場人物 :

内容 : ノードを場面、リンクを時
系列とする有向グラフ。
キーワード

日時、場所、登場人物は抽象化してもよく、それらは概念階層で制御される。QUIK ではたとえば以下のように記述される。

場面 ID/[日時 = X,

場所 = Y,

登場人物 = Z,

内容 = {W₁, W₂, W₃, W₄}]

|{W₁ ≦ W₃, W₂ ≦ W₃, W₃ ≦ W₄}

ここで “ \leq ” は時間的順序を表している。制約部分で、部分場面間の時間的制約を表現している。内容的には文献

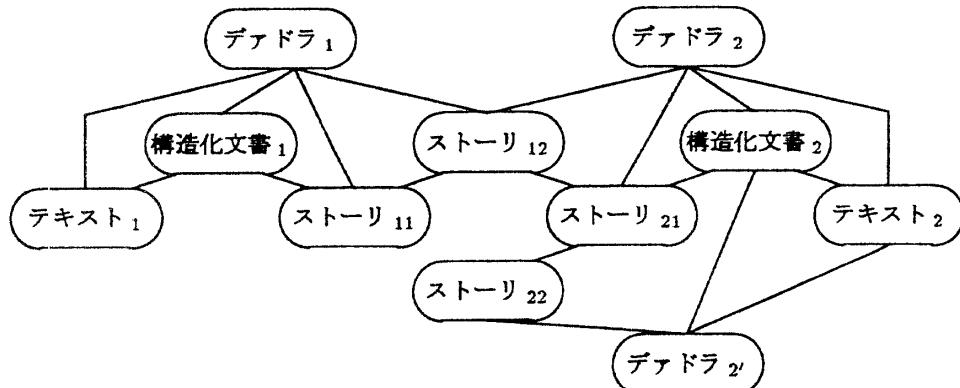


図1: メディエータとしての文学データベース

[3] のように、因果関係など詳細な関係が表現できているのが理想的だが、それは本質的な困難さを伴うので、まず時間的順序で場面間の関係を表現し、次の段階でさらに意味内容を付加することを検討している。

「内容」部分が段階的に詳細化されると共に、人によって複数の解釈に枝分かれする。このような抽象化構造は文献[4]の階層に対応している。それらをテキスト文書や構造化文書と相互に対応づける必要がある。対応付けのための識別子は以下のようにになっている。

- | | |
|--------|---------------------|
| テキスト文書 | 文書の位置 (位置と長さ) |
| 構造化文書 | 構造識別子 (章、パラグラフ、文、…) |
| ストーリ記述 | 場面ID |

4 文学データベースとメディエータ

3節のように構成される文学データベースを図1のようにメディエータシステムとして配置することを考えている。ここで文書の各抽象化階層がそれぞれ1つのメディエータに対応している。それぞれの識別子間の対応は上位のメディエータ (デアドラ_X) で保持されている。ここでストーリ₁₂はデアドラ₁とデアドラ₂で共有化されており、ストーリ₂₂はストーリ₂₁に対する別解釈となっている。

5 おわりに

本稿では知識表現言語 QUIK に基づくメディエータシステム上での文学データベースの構成について議論した。現在デアドラ伝説の電子化と共に、内容に記述実験を行なっている。また QUIK システムは現在 Java で実装中で、年内には文学データベースのプロトタイプを作成予定である。

今後の課題として、専門家による文学データベースの評価と QUIK へのフィードバックを行なうことを考えている。

謝辞

種々の議論を頂きました横田研究室の皆様および岡山理科大学劉助教授に感謝致します。なお、本研究は文部省科学研究費（重点領域研究(1)）によるものである。

参考文献

- [1] 萬上裕、黒田崇、横田一正，“分散環境における仮説生成による問合せ機能の拡張”，情報処理学会、データベースシステム研究会，神戸，Jan. 21-22, 1996.
- [2] 三宅忠明，“ケルトの神話と伝説 — 悲恋ロマンス、デアドラについて”，岡山ケルト文化研究会，Mar., 1996.
- [3] Kuniaki Uehara, Meguru Oe, Keita Maehara (Kobe U.; Japan) “Knowledge Representation, Concept Acquisition and Retrieval of Video Data”, Proc. Int. Symp. on Cooperative Database Systems for Advanced Applications, Kyoto, Dec, 1996.
- [4] 横田一正、柴崎真人，“データベースに判決は予測できるか？”，情報処理学会データベースシステム研究会 & 電子情報通信学会データ工学研究会合同ワークショップ (EDWIN)，長崎，Jul., 1993.
- [5] Kazumasa Yokota, Yutaka Banjou, Takashi Kuroda, and Takeo Kunishima, “Extensions of Query Processing Facilities in Mediator Systems”, Proc. International Workshop on Knowledge Representation Meets Databases (KRDB'97), Greece, Aug. 30,, 1997.