テキストベースの一提案

5Y－7

徳永秀和 青江順一
高松工業高等専門学校　徳島大学工学部

1．はじめに

現在作成される文書は、ほとんどがワープロなどを使用して作成され、コンピュータ上に電子化された形で保存されている。その形式はワープロ、TEX、SGML、HTMLなどである。次にこれらの電子化された文書の再利用のためのツールの現状を考察してみる。アプリケーションとしては、文献データベース、新聞記事検索、ホームページ検索とインデックスサービス、オンラインヘルプ、グループウエアの文書管理などが活用されている。あらかじめ登録されているか、全文検索するかの違いはあるが、検索手法の基本はキーワード検索である。HTMLとオンラインヘルプについては、これにリンク機能が追加されている。また、文書の利用単位は文献単位から節単位であり、検索にかぎれば文献単位（ホームページ）である。

このような現状の文書利用では、電子化された文書を人間の知識保存の道具というにはあまりにも不十分である。そこで本論文では、知識保存の道具として十分に利用できるものとしてテキストベースを提案する。

2．テキストベースに求められる機能

各個人にとって文書を知識保存の道具とするために、提案するテキストベースには、以下のような機能を持たせる。①自分の現有知識を理解してもらえること。②現在の検索目的を理解してもらえること。③複数の文書から検索目的に合った文書を生成できること。④一気に読める程度の回答をもらい、それを読むことにより新たな検索目的を持てること。①と②は、検索履歴の処理とユーザインタフェースの設計という問題となる。③と④は、文書のデータ構造と探索の問題となる。 文献や記事単位での探索に対しては、①、②の研究は行われているが、複数の文書を統合したより細かい単位での探索に対しては行われていない。本研究では、複数文書をいかに統合し、探索に適したデータ構造にするかが第一のポイントである。したがって、以下でテキストベースのためのデータ構造についての提案を行う。ただし、対象とする文書は技術解説文書にかぎり、小説や一般的な新聞記事などは考えないものとする。

3．文書管理の基本単位

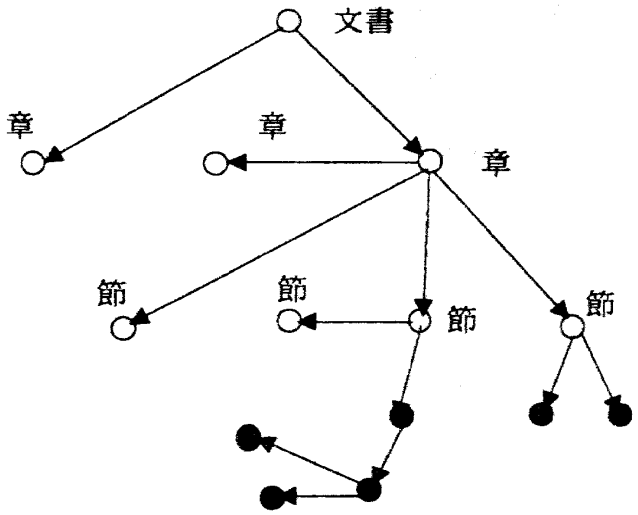
文書管理の基本単位は、数個の文程度とする。そして、この文章で説明したい用語を主要語として1個と、主要語を説明するために使用する用語を関連語として数個もっているものとする。この主要語や関連語は、形容詞や副詞などの修飾語を含むことを許した専門用語または固有名詞とする。この基本単位をテキストアトムと呼ぶことにする。ただし、主要語も関連語も持たないテキストアトムも存在するものとする。

4．単独文書の構造

まず、既存の文書の構造を解析することから始める。既存の文書は、特定の知識を有する読者を対象に、ある分野のある範囲の内容を、ある角度で理解してもらうことを目的に書かれている。そして、基本的には最初から最後まで一本の道としてまっすぐに読んでもらう事を期待している。また、区切りとして章や節を設けて

---

a proposal for the TextBase
Hidekazu Tokunaga, Junichi Aoe
Takamatsu National College of Technology
Tokushima University



● テキストアトム

図１ 単独文書のグラフ

いる。これをグラフとして表現すると図１のようになる。節を１つの点とする。節内のテキストアトムは、節を始点とした半順序の連接木で構成する。そして、章は節の始点として、文書は章の始点として表現する。そして、枝に事例や詳細などの意味を付加する。特殊なものとして章の先頭に要約文などがある。これは、孤立点とし、内部の点からは関連する節または節内の点への枝を付ける。以上の構造は、著者が意図した特定の視点からの説明のための単純なグラフである。

しかし、同じものを様々な始点から説明した多くの文書が存在する。そして読者は、著者の視点とは少しずれた視点で情報を得たい場合が多い。したがって、多数の文書をうまく結合したグラフを作成し、読者の視点に沿った文書を提示するシステムが求められる。次に文書間の結合方法を考察する。

５．文書間の結合

まず、文書間の親和度を導入する。文書と章の題名中の主要語、文書と章の要約文中の主要語のマッチング度合いによっ文書間の親和度を決定する[1]。そして、親和度の値によって、章間までの枝を追加するか節間までの枝を追加するかを決定する。親和度の低いものは、テキストアトムと節間とテキストアトム間の枝の追加にとどめる。

６．追加する枝

章間や章と節間に追加する枝は以下のようなものを考える。①ほとんど同じ事を記述している。②どちらかが他方の事例や詳細を記述している。③同じレベルで他の製品や技術の話をしている。①、②は、章や節の題名、内部のテキストアトムの主要語から判断が可能[2]。③は、文書構造と章、節の題名から判断する。テキストアトムへの枝は、同じ主要語同士、関連語と主要語が同じものの間に付ける。

７．適応実験

日経オープンシステムの特集記事より、イントラネット関係の記事３件に対して、記事間の枝を追加する実験を行った。親和度は、題名や主要語イントラネットやWWWが多く存在するために非常に高い。そこで、章、節間の枝について検討すると、①、②はいくつか見つけることができた。 ただし、枝の決定演算は今後の検討課題であり、今回はヒューリスティックに行った。最後に具体的な数を示しておく。章、節の数は４２個である。①に対応するものが７件、②に対応するものが６件であった。

８．おわりに

複数の文書をグラフとして表現するための案を今回提案することができた。今後は、１年分程度の雑誌の技術解説記事に対して適応し、質問の仕方およびそれに対する提示文書の探索方法を検討していく。

参考文献

[1] 大竹清敬、増山繁、山本和英 ”名詞を中心とした連接に着目した新聞の関連記事検索手法” 情報処理学会研究報告 97-NL-122-12

[2]柴田昇吾、上田隆也、池田裕治 ”複数文書の融合” 情報処理学会研究報告 97-NL-120-12