シソーラス関連性を利用した異文書群の共通SGML文書構造化

藤岡 孝子*

大日本印刷株式会社 C&I総合企画開発本部 C&I研究所

1. はじめに

SGML文書はXML文書として新規応用の可能性が高まっているが[1]、全文DBなどへ実用化される文書群はフォーマットが決まった特定分野,単一目的の大量文書が多い[2]。

しかし日常的には、大量の紙面情報が決まった索引管理もされずに生産・出版されており、我々はこのような文書の構造化する試みを進めている[3]。文書が単一目的で使用される場合は構造定義を特定の形態にあわせればよいが、通常の雑誌など「眺めて楽しめる情報」は、内容的に一紙面中に多様な観点の情報を含み、ユーザの自由な視点で読めるよう編集されているため、電子文書とは違って簡便に多様な情報を入手できる情報源として利用されている。

一般の電子文書の場合、ある主題についてキーワード検索すると、ばらばらな視点で検索されてしまい、多様性から生まれる面白味を楽しむことはできない。また、単一情報の条件検索では、ユーザの自由な興味の移動ができず、やはり「眺めて楽しむ」面白さは生まれないと言える。

そこで我々は、互いに関連する主題をもつ異文書群をとりあげ、多角的な視点で同時に内容情報を提供できるようなコンテント構造の実現を目指し、シソーラス関連性を利用した共通SGML文書構造化を試みている。

本稿では、まず、どのような異文書群を扱うかに触れ、次に、文書の主題キーワードと視点の違いについて述べる。さらに、シソーラスを用いて共通SGMLタグ定義を作成する方法を説明したのち、構築中の実験システムの概要について紹介する。

2. 異文書群の共通統合化

2.1 単一目的型文書と複合型文書の違い

我々のターゲットは、多様な情報を同時にユーザに提供できる文書コンテントの構造の作成である。このために、まず、単一目的型文書構造と、複合型文書構造について考察する。

1) 単一目的型文書

(例：レストランデータブック、料理手順書)

・同一の形態にそろえられた部分的な情報が並び、読者の文書を読む動機がはっきりしている。

・個々の情報の項目がはっきりしており、条件に応じた索引がつき、含まれる画像の種類もかなり統一されている。

2) 複合型文書

(例：食文化に関する雑誌、タウン情報誌)

・部分的な情報で構成されるが、均質ではなく、多様な視点の情報が効率よく収められている。

・ユーザ（読者）に、興味や目的の移動を自由に許すような構造になっている。

以下に例を示す。主題は「バジリコのスパゲティ」だが、部分的にはレストランガイド情報、料理手順書、食材辞典、などの文書が含まれており、レイアウト上、ユーザが注目した部分の近傍に、詳細テキストが細かい体裁で配置される。

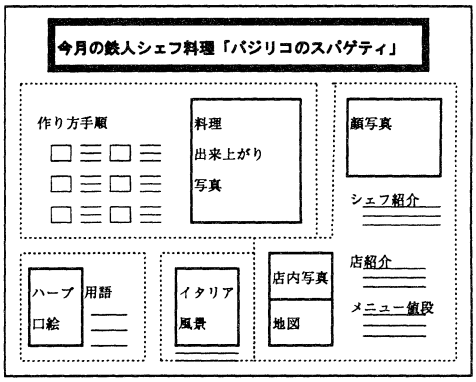

図1. 複合型文書のレイアウト構造例

2.2 文書中の主要キーワードへの視点情報

上図例に示したような文書紙面を電子的に実現するには、複数の単一目的型文書群から構成でき、またこのような紙面から単一目的の部分構造を取り出せるような文書の目的を性格づける特徴が必要である。

そこで、各文書群に共通して出現する共通主題キーワードと、それへの関連視点情報を定義することにより単一目的型文書を性格づけ、さらに互いの文書を関連付けることを考える。すなわち、文書中の重要語の性質を以下の3つに分類し、蓄積する。

1) 単一型文書のキーワード

単一型文書中、索引、見出しにある名詞、出現頻度の高い名詞と、シソーラス上の類語グループ

(例) 料理書→作り方、食材名、料理名

レストラン情報→飲む、飲食店分類名

2) 共通主題キーワード

上記のキーワードで異文書群で共通するもの。

3) 関連視点情報

文書が読者のどんな興味に応えようとしているかの情報であり、書名、前記の主要キーワードの近辺の見出しや索引に頻出する語句。

* 連絡先：藤岡 孝子　大日本印刷（株） C&I総合企画開発本 C&I研究所 〒162-0066 新宿区市谷台町6-10
TEL: 03-5269-5456 FAX: 03-5269-5461 E-mail:fujioka@lab.cio.dnp.co,jp

（例）値段、営業時間、利用する、作る

1）では文書中出現パタンが共通している抽出語のうち、頻度の高いシソーラス上のグループを選択し、その上位概念語か、文書情報のレイアウトからわかるものでラベルづけして蓄積する。各文書群からグループができた後、共通のグループがあれば主題とする。

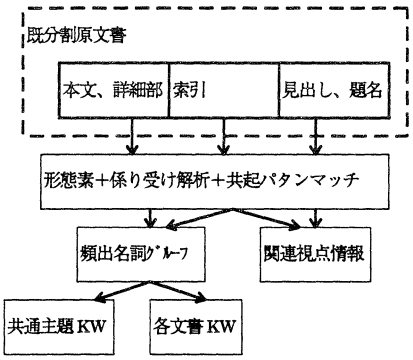

図2 共通主題 KWｸﾞﾙｰﾌﾟと関連視点情報の作成

3. シソーラスを利用した共通 SGML タグ作成

互いに異なる目的の文書に共通して使えるタグとして、以下のようにタグ構造を作成する。タグをもつキーワードのデータ自身もSGMLファイルとして作成する。この手順を以下に説明する。

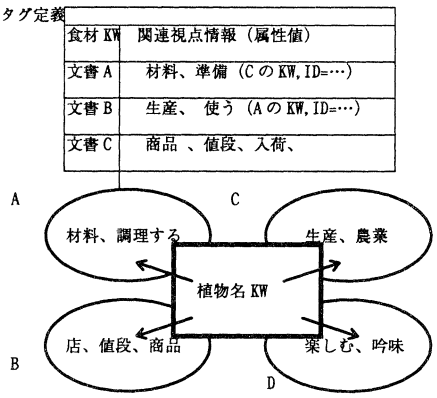

図3．タグ定義の作成

1）共通主題キーワードグループには EDR 上の上位概念語の意味 ID からタグ名を与え、下位分類を辞書より作成し、下位概念語をその下に登録する。

2）共通主題および各文書キーワードのタグに対して、原文書から得られた視点情報を原文書の分類と値の組を属性として登録する。

3）主題と視点が、他の文書の固有キーワードグループと関連する場合はそのキーワードグループを示す意味 ID を値に与える。

4）各文書中の固有キーワードが、何らかの実際の「店」や「人」などを指していることがあれば、そのキーワードへの「実体情報」属性として名前や年齢、住所などをさらに属性情報として与え、条件により文書中から探せるようにしておく。

4. デモシステム概要

文書サンプルデータは角川書店「TokyoWalker」および NHK 出版「きょうの料理 CD−ROM」を用いて構成した。ユーザがある文書情報に注目していても、それのみを表示せず、関連する多様な情報を常に保持するようなブラウジングができるようにしている。

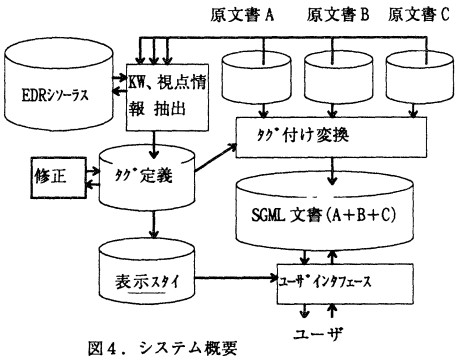

図4．システム概要

まとめ

文書ごとに異なる主題への視点情報を取り入れ、共通する主題キーワードに関する異文書群を共通 SGML 化し、ユーザへの多様な情報提供手段としての実験システムを構築しデモを行った。「眺めて楽しい」電子文書実現のために、今後も内容情報への索引付けの多様化と提示方法への工夫を試みると同時に、XML 対応型文書配信方法への展開も考えていきたい。

謝辞

「TokyoWalker」データの使用を許諾してくださいました角川書店雑誌事業部、ならびに「CD-ROM 版 NHK きょうの料理大百科」の使用を許諾してくださいました日本放送出版協会および NHK エデュケーショナルに感謝いたします。

参考文献

[1]W3C,http://ww.w3.org/XML/

[2]成田,松本:論文全文データベース作成のための SGMLﾀｸﾞ付けの自動化,言語処理学会第 1 回年次大会論文集,pp.329-332,1995

[3]藤岡:文書中の重要語の意味的分類に基づく SGML 文書構造化,人工知能学会全国大会,1998

[4]EDR 電子化辞書 v1.5,日本電子化辞書研究所,1995