XML文書を対象とした例示検索法の検討

5V－4

長谷川 知洋　梅田 昌義　谷口 展郎　山室 雅司
NTT情報通信研究所

1. はじめに

インターネットの爆発的な普及によりネットワーク上のみならず、オフィスなどのローカルな環境においても大量の電子化文書が氾濫してきた。電子化文書の中でも構造化文書は広範囲なアプリケーションで利用され、その数は急激に増加しつつある。それに伴い、大量の文書情報の中から必要なものだけを検索したいという要求が高まってきている。

XML[1]は WWW 上での利用を目的とした文書記述言語であり、簡易版 SGML として普及が期待されている。XML の普及により今後益々、構造化文書が利用されることが予想される。

本稿では、XML 文書を同一文書中にテキスト、画像、音声など複数のメディアを格納した複合メディア文書として捉え、希望通りの検索結果が簡単に得られるような類似検索法を提案する。

2. 複合メディア文書の検索

2.1. 複合メディア文書

XML 文書はテキスト情報に加え、画像や音声といった複数のメディアを格納することが可能である。そこで、このような文書のことを**複合メディア文書**と呼ぶ(図 1)。複合メディア文書のアプリケーション例として、Web ページ、電子図書館、電子カルテなどが考えられる。

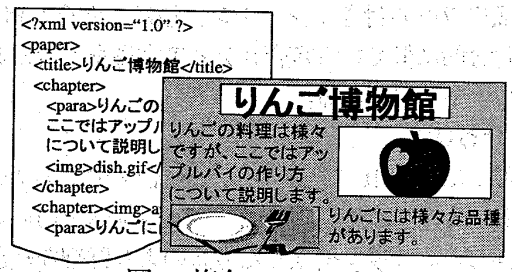

図 1 複合メディア文書

2.2. 関連研究

複合メディア文書を対象とした検索を考えた場合、従来 Web 検索などで用いられてきた技術は、キーワード入力による全文検索がほとんどで検索結果にノイズを多く含んでしまうという問題がある。一方、[2]によると構造化文書の検索では、主にテキスト情報を管理する索引と文書の論理構造情報を管理する索引を用いた一致検索や範囲検索を行うことによりノイズを削減することができる。しかし、検索キーの一部に構造情報も与えるためには、検索者が文書構造(タグ名など)をあらかじめ知っていなければならない。また、複合メディア文書に含まれる画像などの有力な手掛かり情報を検索キーとして使用することができない。

2.3. 複合メディア文書を対象とした類似検索法

本稿では、検索キーの入力が容易な例示による類似検索法(例示検索法)を提案する。類似検索の一般的な方法は、まず検索対象のデータからデータの特徴を示す情報(特徴情報と呼ぶ)を抽出し、数値化する。そして、数値化した特徴情報(特徴量と呼ぶ)を比較することで類似性を判定する。我々は、この方法を複合メディア文書に適用することを検討している。

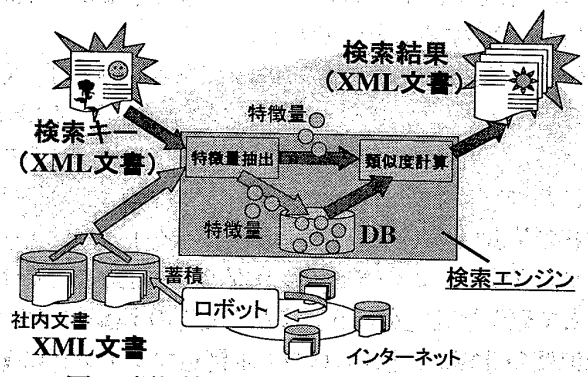

図 2 類似検索システムの実現イメージ

提案する方法では、複合メディア文書自身を検索キーとして例示し、例示された文書に含まれるテキスト情報、画像情報、論理構造やレイアウトの情報に基づき、類似した文書を検索する(図 2)。この特徴情報として、以下のようなものが考えられる。

[内容の特徴情報]
- ◆ テキスト(記述内容、単語の順序関係や出現頻度)
- ◆ 画像(色相、彩度、輝度、色配置)

[構造の特徴情報]
- ◆ エレメント名(タグ名のこと)、属性情報
- ◆ 文書の木構造(形状、高さ、ノード数)

[レイアウトの特徴情報]
- ◆ テキスト(位置、領域の割合)
- ◆ 画像(位置、大きさ、領域の割合)

上記システムに基づく問合せ例を以下に示す。例えば、問合せ例aは内容の特徴情報を、問合せ例bは構造の特

A Study on Query by Example for XML Documents
Tomohiro HASEGAWA, Masayoshi UMEDA,
Noburou TANIGUCHI, Masashi YAMAMURO
NTT Information and Communication Systems Labs.
1-1 HikarinookaYokosuka-shi Kanagawa 239-0847

徴情報を利用することで検索可能である。

[問合せ例]

a. この文書に類似した内容の文書を探せ
b. この文書の定型フォーマットと似たフォーマットを持つ文書を探せ
c. この文書とレイアウトが似ている文書を探せ
d. この文書とそっくりな文書を探せ

3. 複合メディア文書の類似性判定法

現在、複合メディア文書の類似性を判定するのに適した方式が存在しないため新たに二つの方式を提案する。

3.1. 重み付き線形計算法

これは文書に含まれるテキスト情報や画像情報、文書構造やレイアウト情報の類似度を個別に計算し、それらに重みを付けて線形和をとったものを複合メディア文書としての類似度とする方式である。重みを適宜変更することで、2.3節で示したそれぞれの問合せ例に対応可能である。例えば、テキストや画像の類似度に対する重みを強くすることで問合せ例aに対応することができ、文書の木構造の類似度に対する重みを強くすることで問合せ例bに対応することができる。

個々の類似度の計算方法は様々である。例えば、文書中のテキストや画像など、主に文書内容に関する類似度の計算は、[3]で採用している多次元ベクトル空間モデルに基づいて行うことができる。これは2.3節で挙げた内容の特徴情報を数値化し、多次元ベクトル空間上へマッピングする。このベクトル空間上での距離を計算することで類似性判定を行う。また、文書の木構造に関する類似度の計算は、[4]に基づく木構造マッチングに基づいて行うことができる。

3.2. 属性付き順序ラベル付き木マッチング法

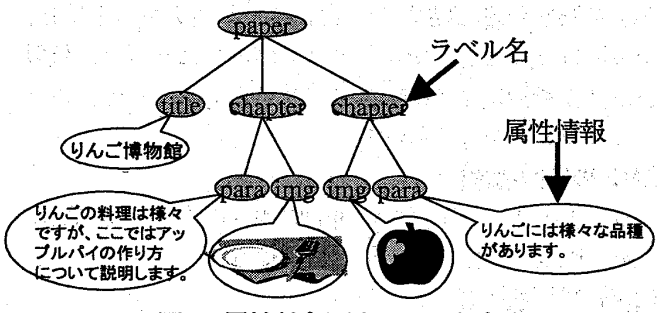

図 3 属性付き順序ラベル付き木

これは[4]に基づく木構造マッチングを行う際に木の形状だけでなく、木のノードに付加されているラベル名やノードがもつ属性情報を考慮して類似性判定を行う方式である。文書構造は、エレメント名をラベル名とする順序ラベル付き木(ordered labeled tree)として表現できることが知られているが、エレメントにマークアップされたテキストや画像などの内容情報は、ノードがもつ属性情報と考えることができるので**属性付き順序ラベル付き木**(図3)と見なすことができる。

3.1節(重み付き線形計算法)での木構造の類似度が、順序ラベル付き木の形状だけから計算されるのに対し、本方式では、木のノードに含まれるテキストや画像などの内容を考慮して木構造の類似度(これが複合メディア文書の類似度となる)を計算するため、適合率の向上が期待できる。

3.3. 考察

3.1節及び3.2節で提案した両方式のモデル図を図4に示し、両者の比較を表1にまとめる。

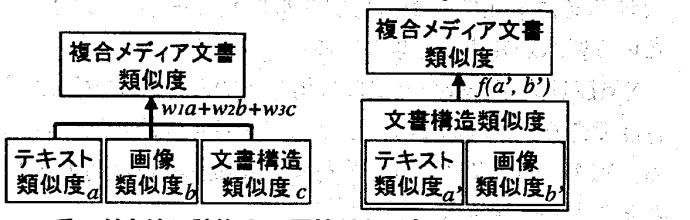

図 4 提案方式のモデル図

表1 類似性判定方式の比較

比較項目	適合率	計算量	検索の柔軟性(重み調整)
重み付き線形計算法	△(重み調整に依存)	○	○
属性付き順序ラベル付き木マッチング法	◎	△	△

両者の優劣については、適用する問合せの種類によって変わってくる。例えば、2.3節で示した問合せ例a～cは、重み付き線形計算法が適していると思われる。属性付き順序ラベル付き木マッチング法は、問合せ例dのように総合的に似ている複合メディア文書を検索するのに適していると思われる。

4. おわりに

本稿では、XML 文書のような複合メディア文書を対象とした類似検索の一手法として例示検索を用いることを提案した。今後は提案した方式のより詳細な検討を進めていく。

参考文献

[1] W3C, "Extensible Markup Language(XML) 1.0," http://www.w3.org/TR/REC-xml, 1998.
[2] 金本, 加藤, 絹谷, 吉川, "効率的な更新が可能な構造化文書の索引," DBS 114-10 pp65-72, 1998.
[3] K. Curtis, N. Taniguchi, J. Nakagawa and M. Yamamuro, "A Comprehensive Image Similarity Retrieval System that utilizes Multiple Feature Vectors in High Dimensional Space," DBS 113-28 pp167-172, 1997.
[4] J. T. L. Wang, K. Zhang, K. Jeong and D. Shasha, "A System for Approximate Tree Matching," IEEE Transactions on Knowledge and Data Engineering, 6(4), pp559-571, 1994.