6W-4

複数のメディアで構成されたXHTML文書の検索手法

鈴木 優† 波多野 賢治† 吉川 正俊†‡ 植村 俊亮†

† 奈良先端科学技術大学院大学 情報科学研究科

‡ 国立情報学研究所 ソフトウェア研究系

概要

本稿ではXHTML文書に含まれる文字列，画像，文書構造を考慮して検索する手法を提案する．従来の構造化文書検索の研究では文字列と文書の木構造が検索対象とされていたが，画像や映像を扱うことのできる構造化文書を検索することができない．本研究では複数のメディアを扱うことにより検索精度を向上させる方法を提案する．

1 はじめに

現在，WWW (World Wide Web) の普及によりHTML (Hypertext Markup Language) が様々な場面で用いられている．一方，異なるシステム間のデータを交換するためのフォーマットとしてXML (Extensible Markup Language) が用いられている．HTMLは文書を表示させるための言語であるのに対し，XMLはデータの再利用を念頭に置いた言語である．また，これら二つの特徴を組み合わせたXHTMLが提案された．このため，将来的にはXHTMLで記述された電子文書が今後広く用いられると考えられる．

HTMLやXHTMLはテキストだけでなく画像や映像，音声などを扱うことができるフォーマットであるが，現在のWeb検索エンジンではそれらを統合した検索ができない点は問題である．つまり，現在のWeb検索エンジンでは例えば「"横浜"というキーワードが含まれる文書」のように単語の羅列による問合せが主に使われている．ところが，以前一度見たことのあるWeb文書を検索したい場合には，キーワードを入力するよりも「黒っぽい画像が右上にあるもの」のように画像やそのレイアウトなどを問合せとして入力したほうが，より利用者に興味のある電子文書を検索することができると考えられる．著者らの以前の研究 [1] ではPDF文書に対してこれらの問合せを実現したが，本稿ではXHTMLを対象としてテキストや画像，またそれらの構造を問合せとして利用するための方法について論じる．

A Retrieval Method of XHTML Documents Consist of Multimedia Data

Yu Suzuki†, Kenji Hatano†, Masatoshi Yoshikawa†‡, Shunsuke Uemura†

† Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

‡ Software Research Division, National Institute of Informatics (NII)

2 XHTMLの検索手法

本手法では，複数のメディアからなる電子文書から複数の特徴量を抽出し，それらをベクトル表現する．問合せも同様に問合せ拡張を行うことによってベクトル表現し，それぞれのベクトルの近接を求めることによりオブジェクトの評価値を得ることができる．本稿でXHTML文書から取り出す特徴量を以下に示す．

- **テキスト，画像の特徴量**
- **文書構造の特徴量**
  XHTMLなどXMLに基づいて記述される文書の構造は，木構造として表すことができ，その木構造はDOM Tree[2]と呼ばれている．

これらの三つの特徴量のうち，テキストや画像の特徴量のベクトル化については既に以前に多くの手法が提案されている [3, 4] が，文書構造の特徴量をベクトル化して検索する手法については，我々の知る限りまだ提案されていない．そこで，次節で文書の木構造をベクトル表現する方法について述べる．

2.1 XHTML文書の木構造のベクトル化

2.1.1 XHTML文書の木構造による表現

例として図1のXHTML文書を考える．簡単のために，タグ名を *a, b, c* とする．図1を木構造で表現すると，図2のようになる．この例の場合，三つのノードがXHTML文書に含まれている．

2.1.2 ベクトルの基底の作成

次に，この文書の構造ベクトルの基底を求める．まず，文書に含まれている各ノードへの経路式を求める．

```
<?xml version="1.0" ?>
<a>
   <b></b>
   <c></c>
</a>
```

図 1: XHTML 文書の例

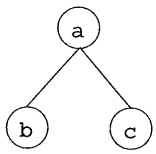

図 2: XHTML 文書の木構造

1. /a (ノード a への経路式)
2. /a/b (ノード b への経路式)
3. /a/c (ノード c への経路式)

次に，これらの経路式の接尾辞を求める．ここで，接尾辞とは文字列の末尾の集合である．

- /b (/a/b の接尾辞)
- /c (/a/c の接尾辞)

以上の方法で求めた経路式とその接尾辞の種類を文書の構造ベクトルの基底として定義する．

2.1.3 文書構造の特徴ベクトルの作成

最後に，文書構造の特徴ベクトルを求める．あるノード t への経路式 /t の出現頻度を $x(/t)$ と定義すると，文書構造の特徴ベクトル $\boldsymbol{X}$ は次のように計算される．

$$\begin{aligned}\boldsymbol{X} &= [x(/a), x(/a/b), x(/a/c), x(/b), x(/c)] \\ &= [\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]\end{aligned}$$

つまり，文書構造の特徴ベクトルを前節で求めた経路式とその接尾辞の出現頻度として表現する．

3 あとがき

本稿では，XHTML から文字列のみではなく画像や文書構造などを問合せとして利用可能な検索手法の提案を行った．本手法を用いる利点を次のことが挙げられる．

- 電子文書の構造をベクトル化し，問合せをベクトル化したものとの類似度を求めることによって，電子文書の構造の適合度で順位付けを行うことができる．
- 利用者が文書を検索する手がかりとして，単語の出現頻度だけではなく文書の構造を利用することで，より利用者の興味を明確に表現することができる．

本研究の課題として次のことが挙げられる．

- ベクトルの要素がそれぞれ相関関係を持っているため，ベクトル演算などを行った場合に正しく計算されない場合がある．例えば /a/b が出現すれば必ず /b が出現するため，これら二つの経路式は独立しているとはいえない．
- タグ名に相関関係があるため，ベクトルの要素が独立しているとはいえない．例えば table タグがあれば必ず tr タグが含まれるため，これら二つの要素は独立しているとはいえない．

謝辞

本研究の一部は，文部科学省科学技術研究費（課題番号: 11480088，12680417，12208032，12780309），ならびに科学技術振興事業団戦略的基礎研究推進事業によるものである．ここに記して謝意を表します．

参考文献

[1] Y. Suzuki, K. Hatano, M. Yoshikawa, and S. Uemura. A Unified Method of Multimedia Documents. In *Proceedings of International Conference on Database Systems for Advanced Applications*, April 2001. (to appear).

[2] L. Wood, et al. Document Object Model (DOM) Level 1 Specification (Second Edition) Version 1.0, Sep. 2000.

[3] G. Salton. *Automatic Text Processing: The Transformational, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1988.

[4] 串間和彦, 赤間浩樹, 紺谷精一, 山室雅司. 色や形状等の表層的特徴量にもとづく画像内容検索技術. 情報処理学会論文誌, Vol. Vol. 40, No. No. SIG3(TOD1), pp. 171 – 184, February 1999.