テキストのXML文書化と全文検索に関する検討

1Z－03

高野哲郎, 佐々木貴文, 上島紳一
（関西大学 総合情報学部 総合情報学科）

1. はじめに

パソコンの普及やインターネット技術の発展に伴い膨大な量の電子化されたテキストが存在する. 代表的なデータ形式として使われているものに XML があり, 今後ますます増加すると考えられる.

本稿では, XML の文書構造を利用した検索手法について検討し, XML テストデータの作成と XML 全文検索システムについて述べる（図１）.

2. 索引の作成と全文検索

2.1 共通パターンを持つ文書のXML化

まず, 特定のパターンを持つ文書の構造を明確化し計算機処理を容易にするため XML 文書に変換する. XML 文書化によりデータの文書構造による意味付けが可能となる.

図２にニュース記事をXML文書化した例を示す. ニュース記事はタイトルや日付・記事のジャンル・記事の内容等の同じ項目の情報が存在することから文書構造が類似しているという特徴

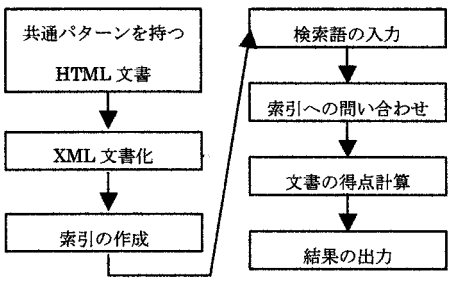

図１ システムの流れ

Generating XML Document and its Full Text Search.
Tetsuro Takano,Takafumi Sasaki,
Shinichi Ueshima.
Faculty of Informatics,Kansai University.

```
<?xml version="1.0" encoding="Shift_JIS" ?>
  <news resource="Yahoo!ニュース"
  company="毎日新聞">
  <genre category="スポーツ">
    <article time="200111171521" title="＜ＮＢＡ＞ジョ
    ーダンが今季最多得点４４点もチームは６連敗">
      <para>
      ＮＢＡのウィザーズで現役復帰したマイケル ジョ
      ーダンは１６日、ジャズ戦で４１分間プレーし、今
      季ＮＢＡ最多タイとなる４４得点を記録した。
      しかし、チームは９２却１０１で敗れ６連敗となっ
      た。　共同 毎日新聞
      </para>
    </article>
  </genre>
</news>
```

図２ ニュース記事をXML文書化した例

を持つ. 記事データの類似する文書構造から推測されるDTDを１つのパターンをして捉え, 記事データをこのDTDに則してXML文書に変換した.

2.2 XML 文書の索引の作成

XML 文書に対する全文検索を行うために, その準備として検索に必要な索引ファイルを作成した[1]. 索引に格納する単語の抽出は日本語形態素解析ソフト「茶筅」を用いている[2].

索引ファイルには各単語を含む文書名, 単語のパス情報（その単語に至るまでの経路）, 単語の文書内の位置情報を格納する.

単語のパス情報とは, 各単語が文書構造上の階層の位置を示すものである. 例えば, 図２の“ウィザーズ”という単語は「ウィザーズ１１１１」と索引に格納されるこれはウィザーズという単語が元の XML データ（図２）のルートから１番目の１番目の１番目の１番目の１番目の子要素に含まれていることを意味する.

2.3 XML 文書に対する得点付け

検索結果とする文書に対しては得点付けを行う. Namazu などの全文検索では TF/IDF 法が主

に利用されているが，ここでXML文書の特徴を

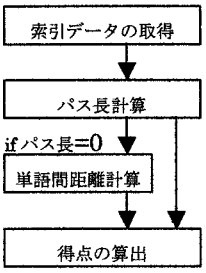

図３ 得点付けの流れ

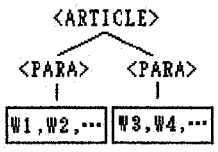

図４ XMLインスタンス

生かした図３のような得点付け方法を提案する.

索引ファイルに記録されているそれぞれの単語のパスを比較する.それにより得られた単語間のパス長に対し重み付けを行い得点とする.パスが完全に一致する場合は単語間距離を利用する.

図４はXMLインスタンスをツリー状で表したものである.例えば「W1」と「W3」の場合，２単語間を結ぶ経路上の枝の数は４本でありパス長は４となる.

検索対象となる単語が文書内で複数ある場合はそれぞれの組み合わせで得点を計算し最も得点の低いものを採用する.このための計算式は次のようになる.W1,W2は検索語i,jは文書内でi,j番目に表れるW1,W2であることを表す.

$$min\{dist(W1i,W2j)\}$$

検索語がｎ語（ｎ≧３）の場合,nC_2通りの組合せについて上記の計算を行い,その総和を得点とする.例えば検索対象語が「W1」「W2」「W3」の場(W1,W2),(W1,W3),(W2,W3)の３通りについて得点を求めその和をその文書に対する得点とする.式は次のようになる.

$$\sum_{m=1}^{I-1}\sum_{n=m+1}^{I} min\{dist(W_{mi},W_{nj})\}$$

本システムではTF/IDF法とは異なり得点の低いものの方が順位は上に表される.

3. 実験結果

「Yahoo!ニュース」の記事とTF/IDFを用いた検索をし,検索結果の比較により本システムの検証を行った.

検索結果は次のように分類した.

検索対象となる文書が

Ⅰ 検索語を主題とする.

Ⅱ 検索語が主題ではないが検索語についての記述が存在する.

Ⅲ 検索語と関連性がない.

「NBA・ジョンソン」など複数の単語で構成される検索語を数通り入力し検索を行ったところ検索結果に次のような特徴が表れた.

・単語の位置情報の差で得点を算出しているものの多くはⅠにあたる.

・得点が大きく特にパス長で得点付けを行っている文書はⅡがほとんどであった.

・TF/IDFによる検索の結果は上記と異なり上位にもⅢに該当する文書が含まれていた.

・Ⅰに該当する文書の中にはXML検索では下位になったがTF/IDFでは上位になっているものがあった.

これらの結果から,単語位置とパス距離のみを考慮した検索を行うと,検索結果として良い結果も得られるが結果として完全ではないことが言える.そのためTF/IDFやXMLタグによる得点の付加等も考慮した検索結果に対する新たな得点付けの方法を今後の課題としたい.

4. おわりに

本システムでは文書構造を利用することで従来のシステムとは違う結果を得ることができた.しかし文書の内容によっては正しく検索されない場合があり更なる改善が必要である.

参考文献

[1]石川佳治,定兼邦彦,北川博之:文書データを対象とした索引技術,情処誌 42 巻 10 号,p.980-987,2001年10月

[2]形態素解析システム 茶筌:
http://chasen.aist-nara.ac.jp/