E-002

XML 電子新書システム

XML electronic pocket-size paperback system

矢島　匡人†　　小池　勇治†　　高野　明彦‡　　絹川　博之†
Yajima Tadahito　　Koike Yuji　Takano Akihiko　Kinukawa Hiroshi

1．はじめに

現在 XML という技術は，既に電子書籍業界にとって標準となっている．この XML を最大限に利用した，高機能な電子書籍が求められている．その一つとして XML 電子書籍システムの開発を進めている．

XML電子書籍システムはXMLを利用することにより，効率よく知識を得ることを目的としている．知識対象となる書籍の種類としてまず以下の３つを想定している．

- 辞典・辞書
- 学術書
- 新書

本研究の XML 電子書籍システムは，これら３つの書籍を相互に閲覧したり鳥瞰したりすることを可能とし，知りたいことに対して平易な説明，学術的な説明，時流に即した説明や事例を得ることを可能にすることを目的としている．XML 電子辞典システム[1]は既に開発が進められている．

XML 電子新書システムは，XML 電子書籍システムのサブシステムであり，新書から新しい知識や情報を効率よく得ることを目的とする．

2．XML 電子新書形式

2．1 新書の論理構造

新書の論理構造として，書名，著者名，出版日，出版社名，ISBN などの書籍情報があり，その他の情報としてカバーや見返しや帯の情報がある．本文やあとがき，まえがきは段落，図や表を含む論理構造体から構成され，この論理構造体は，一般に題名や，番号，を有している．また，論理構造体は下位の論理構造体を有している場合がある．この論理構造体は上の階層から順に章，節，項，と呼ばれるのが普通である．また論理構造体に含まれる図や表には図番号，図題，表番号，表題を持っている．図１で新書の論理構造を示す．

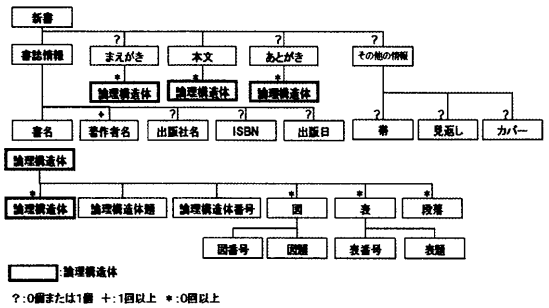

図１．新書の論理構造図

2．2 JepaX 形式による新書表現

新書を XML 新書システムで扱うため，電子書籍業界向けの中立的な形式を指向している JepaX 形式[2]を利用することとし，各論理構造のタグ対応を表１に示す．

表１．新書の情報と JepaX 形式の対応例

新書の情報	JepaX 形式	新書の情報	JepaX 形式
書誌情報	bookinfo	本文	body
書名	book-title	章や節など	div
著者名	book-author	章題や節題など	title
出版日	pub-date	まえがき	front
出版社名	publisher	あとがき	back

2．3 JepaX 形式への変換

岩波形式から JepaX 形式に変換する新書データについて

（1）　書誌情報の要素を抜き出し，JepaX 形式に変換する．JepaX 形式において，岩波形式では表現されていない要素，属性がある場合は新たに作成する．

（2）　新書の内容としてまえがき，あとがきがあるか判別し，ある場合には front 要素，back 要素に変換する．

（3）　本文を JepaX 形式に変換する．

（4）　ファイル名は ISBN で表す．

3．XML 電子新書検索システム

本システムは Web アプリケーションとして開発する．本システムにはユーザの情報要求を満たすための新書を効率よく見つけ出すための機能を作成する．

3．1 書誌情報検索機能

著者や書名などのいわゆる書誌情報を検索対象として検索にかけ，目的の書籍を見つけることを目的とした検索機能を開発する．本機能では著者名，書名，ISBN，出版社，目次情報により検索することができる．また，検索結果について発刊年代により絞り込むことができる．

3．2 XML フラグメント文書・論理構造体検索機能

検索質問と文書の類似度を計算することにより検索結果から目的の文書または論理構造体を得ることを目的とした機能を提案する．索引付け，索引の重み付け，検索質問・文書間の類似度の計算式，検索質問の表現として提案されている XML の構造を考慮する手法[3]を利用する．その中で，索引付けと索引の重み付けについては従来の方式に加え独自の手法を提案する．

従来の検索法では，検索結果としては文書単位でしか得ることができなかった．これを論理構造体単位で検索結果を得る手法を提案する．

3．2．1 XML 文書索引付け

XML の構造を考慮するために索引付けの単位に索引語 t だけではなく，XML の構造 c を入れた(t，c)のペアにより索引付けを行う．論理構造体単位の検索結果を得るために，従来の

†　東京電機大学　大学院　工学研究科
‡　国立情報学研究所

文書単位の索引に加え論理構造体単位の索引を作成する．論理構造体単位の索引は論理構造体を一つの文書とし，索引付けを行うことより得る．

索引付けにおいて論理構造体の階層構造をそのまま表現し索引付けを行うと，検索キーワードがどの階層構造に含まれるかの構造を考慮しなければならない．この考慮を無くすために，索引ペア(t,c)において論理構造体の構造 c は，どの階層も最上位の階層に縮退しているとみなす．このイメージ図を図２に示す．

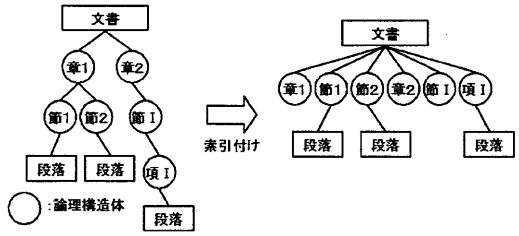

図２.論理構造体索引付けのイメージ図

3. 2. 2 検索質問・文書間類似度計算式

XML の構造を考慮に入れた類似度の計算に，[3]に提案されている拡張ベクトル空間モデルを採用する．

また本機能では，[3]に提案されている構造類似度 cr の測定手法には，構造が一致したときのみに一定の重みを与える Perfect match 手法を用いる．計算式は以下で表される．ここで $\rho(q, d)$は文書 d と検索質問 q との類似度を表し，w_qは検索質問ベクトル(t, c_i)の重みを，w_dは文書ベクトル(t, c_k)の重み，c_iは検索質問索引の構造，c_kは文書索引の構造を表す．$cr(c_i, c_k)$は[3]に提案されている文書検索質問間の構造の類似度を表す．

$$\rho(q,d)=\frac{\sum_{(t,c_i)\in q}\sum_{(t,c_k)\in d} w_q(t,c_i)\times w_d(t,c_k)\times cr(c_i,c_k)}{\|q\|\times\|d\|} \quad (式1)$$

このとき

$$cr(c_i,c_k)=\begin{cases}1 & c_i=c_k\\ 0 & otherwise\end{cases} \quad (式2)$$

3. 2. 3 索引ペア重み付け

重み付けには[3]で提案されている tf-idf・Merge-idf 式とは別に，我々は独自の重み付け構造頻度 icf を提案する．通常，出現回数が増える程，要素の内容が出現回数に比例して増加する．頻出の構造は tf スコアが増え，出現回数が少ないものほど tf スコアが低くなる．このことから出現回数が少ないものほど最終的なスコアが低くなってしまうと考えられる．書名は一つの文書につき一つしか無く，また，論理構造体題も論理構造体の数しか出現しない．これらは重要な情報であるにもかかわらず，頻度が少ないため低いスコアがつけられてしまう．このために構造頻度を重み付けに追加する．構造頻度は全ての構造の出現回数 N を索引ペア(t, c)の構造 c の出現回数 N_cの逆数をとることにより算出する．

tf-idf・Merge-Idf 式は文書の idf スコアを索引語 t と索引構造 c_k の索引ペア(t, c_k)の文書分布ではなく，索引語のみの文書分布により算出する．計算式を以下に示す．ここで tf_dは文書ベクトルの tf スコア，idf は文書ベクトルの idf スコアである．

$$w_d(t,c_k)=tf_d(t,c_k)\times idf(t,C)\times icf(c) \quad (式3)$$

このとき

$$C=\mathrm{Y}_k c_k \qquad cr(c_i,c_k)>0 \quad (式4)$$

$$icf(c)=\log\frac{N}{N_c} \qquad c=c_i=c_k \quad (式5)$$

3. 2. 4 検索質問

要素内容と構造を使用するため，検索質問には，[3]で提案される XML フラグメントを用いる．

4．考察

4. 1 データ増加に伴う索引付け処理の増加

今回提案した，論理構造体単位の索引付けは従来方式に比べ処理が多くなっている．データの増加に伴う，索引付けの処理の増加について今後検討をすすめていく．

4. 2 構造頻度

構造頻度スコアは全体の要素数と相対的に算出している．しかし，この場合データが大規模となった場合，頻出の構造のスコアが非常に低くなってしまうことが考えられる．頻出の構造として，段落要素<p>は検索結果に反映されなくなってしまう．この対策として，XML のスキーマから重みを算出するスコアが考えられる．

5．おわりに

今回我々は XML を利用した，電子新書システムを提案した．その中で，[3]の手法を利用した．さらに我々は以下の３つを提案した．

- 従来の文書単位の検索結果だけではなく論理構造体単位で結果を得ること
- 論理構造体単位で結果を得るために論理構造体単位で索引付けを行うこと．
- 独自の XML の構造にスコアを付与する構造頻度スコア．

今後として，本手法の評価，検討を行っていく．また，他の XML 電子書籍システムのサブシステムとの統合を進めていく．

謝辞

新書コンテンツの提供をいただいた岩波書店に感謝いたします．

6．参考文献

[1] 小池勇治,高野明彦,絹川博之，“複数辞典の鳥瞰が可能な XML 電子辞典システム”,FIT2003 情報科学技術フォーラム情報技術レターズ,pp.95-97

[2] JepaX, http://www. jepax. org/

[3] David Carmel, Yoelle S. Maarek, Matan Mandelbrod, Yosi Mass, Aya Soffer, “Searching XML Documents via XML Fragments”, In Proceedings of SIGIR'03,pp.151-158