XMLデータベースの自然言語検索技術

Natural Language Information Retrieval for XML Database System

真鍋　俊彦　　國分　智晴
■ MANABE Toshihiko　■ KOKUBU Tomoharu

XML (Extensible Markup Language) データをそのまま格納でき，その上で柔軟な検索処理を行えるXMLデータベースシステムの研究・開発を行っている。このシステムの機能強化の一環として，自然文の検索質問に関連する順にXML文書をランキングする自然言語検索技術を開発した。検索質問に関連する語を自動的に追加して幅広く検索する機能や，検索要求に基づいて文書を要約する機能もその一部として実現した。

この技術はXMLデータベースの問合せ言語を通して実行されるようになっており，論理式を基本とした従来のデータ検索や文字列照合を基本とした全文検索と，自然文によるランキング型の文書検索を組み合わせて利用できるようになった。

Toshiba has been developing an extensible markup language (XML) database system with flexible search functions. To enhance the search capability of this XML database system, we have newly developed a natural language information retrieval function on the system. XML documents are ranked in descending order by relevance scores in response to a user's natural language query. In addition, both query expansion and query-based document summarization are realized in this function.

This natural language information retrieval function allows users to utilize the query language of the XML database in combination with Boolean search and full-text search.

1　まえがき

近年，インターネットなどを介して大量の情報にアクセスできるようになり，なかでも情報検索は広く利用される技術となっている。インターネットの検索エンジンを利用することで，ユーザーは入力したキーワードを含む情報を全世界のWebページから検索できる。自然言語検索はこのような情報検索技術の一つで，キーワードだけでなく，自然文の検索質問に対しても情報検索を行えるようにするための技術である。

従来のキーワードに基づく全文検索は，入力したキーワードの出現した文書をもれなく検索できるが，必要な情報を検索するためのキーワードを正確に入力しなくてはならない。これに対して自然言語検索は，欲しい情報を文章で記述すれば，その内容に対応する情報を検索することができる。例えば，企業の相談窓口で顧客から寄せられた質問や苦情をそのまま入力すると，関連する事例を容易に検索できるようになる。

東芝は，情報検索の基盤となるデータベースシステムとして，XMLデータベースの研究・開発を進めている[1]。XMLはインターネット上の交換及び蓄積のフォーマットとして標準化されたデータ記述言語で，日付のような定型データと文章で記述された非定型データが混在する，多様な形式の文書情報を記述することができる。XMLデータベースは，XMLで記述された文書情報であるXML文書を，事前にその構造を定義することなく格納及び管理することができる。このXML文書向けの検索機能強化の一環として，XMLデータベースの自然言語検索技術を開発した。

ここでは，当社が開発した自然言語検索技術とXMLデータベース上での実現方式について述べる。

2　実現した機能

2.1　問合せ言語への組込み

XMLデータベース上で実現した自然言語検索は，自然文で表現された検索質問に対して，その内容との関連度（検索スコア）順に文書をランキングする機能である。この機能をXMLデータベースの問合せ言語であるXQuery（XQuery 1.0：An XML Query Language)[2]から呼び出せるようにした。XQueryはW3C（World Wide Web Consortium）で策定されており，XMLデータベースの問合せ言語として業界標準になりつつある。

現在のXQueryでは，テキストに関する検索条件としては，完全一致検索と前方，中間，及び後方の各部分一致といった文字列照合レベルの仕様は定義されているが，検索結果のランキングには対応していない。しかし，文書検索機能の強化についてはW3Cで議論がなされており，XQuery 1.0 and Xpath 2.0 Full-text（以下，XQFTと略記)[3]として策定が進められている。自然言語検索をXQueryから利用できるように，このXQFTのオプションの一つとして，自然言語検索用のものを当社で独自に定義した。XQFTを拡張した形式で自然言語

```
for $x score $s in db(˝patent˝)/jp-official-gazette
[.//text() ftcontains ˝無線通信における暗号化技術˝
with NLIR]
order by $s descending
return ($s,
        $x//classification-ipc/main-clsf/text(),
        $x//invention-title/text())
```

```
43.181782  H04L9/00  無線通信装置とその通信制御方法
38.209685  H04L9/06  無線通信システムの通信方法
38.204512  H04L9/06  無線通信システム
```

図1. XML文書による自然言語検索の例 — 自然言語で表現された検索質問とその検索結果の例である。

Example of natural language information retrieval of XML documents

検索を呼び出しているXQueryとその検索結果の例を**図1**に示す。

図1は，"無線通信における暗号化技術"が自然文で表現された検索質問で，これを基に特許明細書を検索したXQueryの例である。自然文はこのような短いものに限定されず，例えば，特許の請求項をそのまま入力してもよい。XQueryの3行目に記述された"with NLIR"が，自然言語検索を呼び出すために独自に定義したオプションである。for文の"score $s"という記述により検索スコアが変数$sに格納され，"order by $s descending"という記述により検索スコア$sの降順に検索結果がソート（整列）される。return以下で，検索結果として検索スコア（$s），国際特許分類IPCの筆頭（classification-ipc/main-clsf），及び発明の名称（invention-title）を出力するように指定している。

XQueryから自然言語検索を呼び出せるように実装したことで，特定の構造の文書に限定されずに，以下の利便性をユーザーに提供できる。

(1) XQueryのほかの検索機能，例えば，日付のような定型データ部分を対象にした検索式と，自然言語検索を組み合わせて利用できる。

(2) 文書全体だけでなく，文書の部分構造に的を絞った自然言語検索を実行できる。

上記の(1)と(2)の両方を利用したXQueryの例を**図2**に示す。図2のXQueryでは，検索対象の特許を"国際特許分類IPCの筆頭がG06Fである"と，"出願人に東芝が含まれる"の二つの条件で絞り込まれた文書集合を，"無線通信における暗号化技術"に関して検索した結果を出力する。また，2行目の"(invention-title | claims | abstract)"という記述で自然言語検索の対象を，発明の名称（invention-title），請求範囲（claims），及び要約（abstract）に限定している。

2.2 検索質問拡張

検索質問中の語に対して同義語や関連語を追加して検索する，検索質問拡張と呼ばれる手法がある。XMLデータベース上の自然言語検索では，次の2種類の検索質問拡張を実現した。

```
for $x score $s in db(˝patent˝)/jp-official-gazette
                                        自然言語検索
[.//(invention-title|claims|abstract)//text()ftcontains
˝無線通信における暗号化技術˝
with NLIR]
where
                                        定型データの検索式
$x[contains(.//classification-ipc/main-clsf//text(), ˝G06F˝)]
and $x[contains(.//inventors//text(), ˝東芝˝)]
order by $s descending
return ($s,
        $x//classification-ipc/main-clsf/text(),
        $x//invention-title/text())
```

図2. 他の検索条件と組み合わされた自然言語検索 — 自然言語検索と定型データの検索式を組み合わせた検索例である。

Natural language information retrieval in combination with other types of search conditions

```
for $x score $s in db(˝patent˝)/jp-official-gazette
[.//text() ftcontains ˝検索質問の自動拡張˝
with NLIR [with thesaurus at ˝dic˝]]
                  検索用辞書 ˝dic˝の指定
return ($s,
        $x//classification-ipc/main-clsf/text(),
        $x//invention-title/text())
```

図3. 検索用辞書を利用した検索質問拡張 — 検索用辞書"dic"を用いて検索質問を同義語展開するためのXQueryの例である。

Query expansion using thesaurus terms

(1) 事前に作成された検索用辞書を参照して，検索質問中の語の同義語展開を行う。

(2) 検索処理の中で検索対象の文書から関連語を自動抽出し，検索質問に追加する。

検索用辞書を参照した自然言語検索の例を**図3**に示す。図3のXQueryでは，検索用辞書"dic"を参照して"検索質問の自動拡張"中の語の同義語展開を行う。検索用辞書"dic"の中に，例えば，"検索質問"の同義語として"クエリ"と"検索要求"，及び"問合せ"が定義されていれば，それらの語が検索質問に追加されて検索が実行される。

検索対象の文書から関連語を自動抽出するうえで，擬似適合フィードバックと呼ばれる手法を採用した。擬似適合フィードバックは，検索用辞書をあらかじめ作成するなどの準備をしなくても，検索質問中の語に限定されず幅広く関連文書を検索できる利点がある。XQuery上では，自然言語検索を指定した"with NLIR"オプションの後に，更に"aqe"というオプションを付けると実行される。擬似適合フィードバックの実現方式については3.3節で述べる。

2.3 要約文生成

検索結果のどの文書を参照すればよいかをユーザーが判断することを支援するため，要約文を生成する機能を実現した。要約文は，検索質問中の語が出現した周辺部分の文章を抽出

```
for $x score $s in db(″patent″)/jp-official-gazette
[.//text() ftcontains ″無線通信の暗号化技術″
with NLIR]
order by $s descending
return ($s, $x//classification-ipc/main-clsf/text(),
        $x//invention-title/text(),
        summarize(($x//tech-problem, $x//tech-solution,
                   $x//advantageous-effects, ),
                  ″無線通信の暗号化技術″, 100)
```


43.181782 H04L9/06　無線通信装置とその通信制御方法

無線端末に暗号化のための共通鍵と固有鍵を配置する。固有鍵は各端末に固有のもので，無線端末の台数分だけ定義される。各端末は自端末以外の固有鍵も記憶し，データをマルチキャストする際に，共通鍵と指定された端・・・

100文字の要約文

図4．検索結果の要約文生成 ―"無線通信の暗号化技術"の検索結果について，100文字の要約文を生成した例である。

Query-based document summarization

することで生成される。要約文生成を指定したXQueryの例を**図4**に示す。summarizeが要約文を生成するための関数で，以下の三つの引数を指定する。

(1)　要約文を生成する対象範囲（文書の要素）
(2)　検索質問
(3)　要約文の最大長（文字数）

図4の例では，暗号化技術に関する特許について，検索結果の各文書に関する，課題（tech-problem），課題を解決するための手段（tech-solution），及び発明の効果（advantageous-effects）を対象にした100文字の要約文を生成している。

3　実現方式

3.1　基本的な処理の流れ

当社のXMLデータベースは，次に示す2種類の索引データを利用して，文字列照合を基本とした全文検索機能が実現されている[1]。

(1)　Nグラム索引　　文字列をN文字単位で分割した部分文字列を索引語とし，その出現位置情報を記録した転置索引
(2)　形態素索引　　文字列を形態素解析した結果を索引語とし，その出現位置情報を記録した転置索引

自然言語検索はXMLデータベースに新たな索引データを導入することなく，上記(2)の形態素索引を利用して実現した。そのため，形態素索引を作成しておけば，文字列照合による網羅的な全文検索とランキング型の自然言語検索の両方を実行できるようになっている。

自然言語検索の処理の流れは，以下のとおりである。

(1)　検索質問（自然文）を形態素解析により単語に分割する。
(2)　分割された単語の中から品詞により検索に利用する検索語を選択する。
(3)　検索語と検索対象の文書を照合し，各文書の検索スコアを計算する。
(4)　検索スコア順に文書をソートする。

3.2　検索スコアの計算

検索スコアの計算には様々な方式が提案されているが，情報検索で広く利用されている，確率検索モデルに基づくBM25[4]と呼ばれる方式を採用した。BM25では，検索質問qにおける文書dの検索スコア$S(d)$を以下の(1)式により計算する。

$$S(d)=\sum_{t\in q} tw(t,\ d) \tag{1}$$

ここで，$tw(t,\ d)$は検索語tの文書dにおける重みであり，(2)式で定義される。

$$tw(t,\ d)=\frac{log(|C|/df(t))\times tf(t,\ d)\times(\mathrm{K}+1)}{\mathrm{K}\times((1-\mathrm{b})+(\mathrm{b}\times L(d)\times(\frac{|C|}{\sum_{d\in C} L(d)})))+tf(t,\ d)} \tag{2}$$

C　：検索対象となる文書集合
$df(t)$　：文書集合Cの中で検索語tを含む文書数
$tf(t,\ d)$：文書dにおける検索語tの出現頻度
$L(d)$　：文書dの長さ
K　：$tf(t,\ d)$の影響を調整するための定数（$0\leqq \mathrm{K}$）
b　：$L(d)$の影響を調整するための定数（$0\leqq \mathrm{b}\leqq 1$）

2.1節(2)で述べたように，XQuery上で実行される自然言語検索は，XML文書の全体だけでなく，その部分構造に絞った検索が可能になっている。これに対応するため，(2)式の$df(t)$や$tf(t,\ d)$はあらかじめ用意しておくのではなく，検索処理の中で，XML文書の検索範囲に応じて，索引データ中の索引語の出現位置から計算するようにした。

3.3　擬似適合フィードバック

2.2節で述べたように擬似適合フィードバックは，検索質問に関連語を追加する検索質問拡張の一手法である。擬似適合フィードバックは以下の手順で処理を行う（**図5**）。

(1)　入力された検索質問そのままで文書検索（初期検索）を行う。
(2)　(1)の検索結果の上位文書中の単語について検索語候補としての重みを計算し，その上位一定数の検索語候補を検索質問に追加する。
(3)　(2)で拡張された検索質問で再検索を行う。

(2)の検索語候補の重み$OW(t)$は，以下の(3)式により計算する[4]。

$$OW(t)=rdf(t)\times log\frac{(rdf(t)+0.5)/(|R|-rdf(t)+0.5)}{(df(t)-rdf(t)+0.5)/(|C|-df(t)-|R|+rdf(t)+0.5)} \tag{3}$$

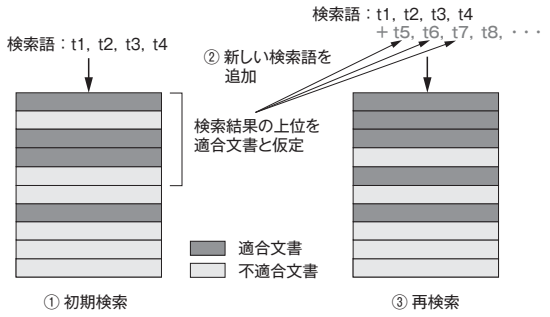

図5. 擬似適合フィードバック ― 検索対象の文書から関連語を自動抽出して幅広く検索することができる。
Pseudo-relevance feedback

C　：検索対象となる文書集合
R　：適合文書と仮定した初期検索結果の上位文書の集合
$df(t)$：文書集合Cの中で検索語tの出現回数
$rdf(t)$：文書集合Rの中で検索語tの出現回数

例えば，"無線通信における暗号化技術"という検索質問に対して擬似適合フィードバックを利用すると，"盗聴"，"送信"，"基地局"，及び"解読"といった関連語が追加される。新聞記事約85万件のテストコレクション[5]による検索実験では，擬似適合フィードバックにより検索精度（平均適合率）が約17％向上するという結果を得た。

4　あとがき

自然文の検索質問で検索を行う自然言語検索をXMLデータベース上で開発した。XMLデータベースの問合せ言語XQueryから呼び出せるようにしたことで，論理式などによる従来のデータ検索と組み合わせて，様々な検索要求に対応できるようになった。また，検索質問に関連する語を追加する検索質問拡張や，検索質問に基づいた要約文を生成する文書要約の機能も実現した。検索質問拡張としては，事前に同義語を定義した検索用辞書を参照する方式と，検索対象の文書から関連する語を自動抽出する擬似適合フィードバックの両方を利用できるようにした。

以上の機能は，東芝ソリューション（株）が商品化し販売しているXMLデータベースTX1$_{TM}$に搭載されている。今後も，文書検索技術の更なる機能向上のため研究・開発を進めていく。

文　献

(1) 宮澤隆幸，ほか．XMLデータベースの全文検索技術，東芝レビュー，**62**，4，2007，p.34－37.
(2) W3C. "XQuery 1.0：An XML Query Language". ＜http://www.w3.org/TR/xquery＞, (accessed 2008-11-14).
(3) W3C. "XQuery 1.0 and XPath 2.0 Full-Text". ＜http://www.w3.org/TR/xquery-full-text/＞, (accessed 2008-11-14).
(4) Robertson, R. E., et al. Simple, Proven Approaches to Text Retrieval. University of Cambridge Technical Report. **356**, 12, 1994, p.1－8.
(5) NTCIR Project. NTCIR-5 CLIR（言語横断検索テストコレクション）. ＜http://research.nii.ac.jp/ntcir/permission/ntcir-5/perm-ja-CLIR.html＞, (accessed 2008-11-14).

真鍋　俊彦　MANABE Toshihiko
研究開発センター　知識メディアラボラトリー主任研究員。
ナレッジマネジメント，情報検索システムの研究・開発に従事。
情報処理学会会員。
Knowledge Media Lab.

國分　智晴　KOKUBU Tomoharu
研究開発センター　知識メディアラボラトリー。
XMLデータベース，情報検索システムの研究・開発に従事。
Knowledge Media Lab.