

ユーザ定義データ型をサポートする XML データの検索モデル

5W-3

石川 佳治[†] 永井 孝明^{††} 北川 博之[†][†] 筑波大学電子・情報工学系 ^{††} 筑波大学システム情報工学研究科

1 研究の背景と目的

XML とユーザ定義データ型 XML は、インターネット上の情報交換のための汎用フォーマットとして大いに着目をあびている。XML に関する最近の取り組みとして、XML 文書構造の定義に対する DTD の能力の貧弱さを克服するための、XML Schema [1] や RELAX [2] などの高レベルの XML スキーマ言語の規格化がある。特に、XML Schema ではユーザ定義のデータ型を導入することを許しており、多様な応用分野において XML を用いた情報の高度な表現を可能としている。一方、スキーマ言語の進展に呼応する形で、XML 問合せ言語 XQuery [3] の提案もなされている。XQuery では、XML Schema で定義されたユーザ定義のデータ型の利用や、ユーザ定義の関数を新たに導入し問合せで用いることが可能である。また、我々の研究グループでもユーザ定義の外部関数を利用可能な拡張可能 XML 問合せ言語 X²QL とその処理系の開発を進めてきた [4]。

地図データの表現と問合せの例 ユーザ定義のデータ型の導入が求められる対象の例として、ここでは G-XML [5] などのように、地図データを XML 文書により表現することを考える。図 1 に文書構造を簡略化した形で示す。GeoSpace 要素は地物データを表す GeoFeature 要素の並びから構成され、GeoFeature は category, Property, Geometry などの要素から構成される。category 要素は、地物データの種別 (例: ビル, 地域) を指示するテキスト要素である。Property 要素は、地物オブジェクトの名前, タイプ (学校, 店など), URL などから構成される。Geometry 要素は、Point, Polygon などの要素として具体化され、地物データの幾何形状を表現する (詳細は省略)。

```
GeoSpace ::= (GeoFeature+)  
GeoFeature ::= (category, Property, Geometry, ...)  
Property ::= (name, type, url, ...)  
Geometry ::= (Point | Polygon | ...)
```

図 1: 地図データ表現のための文書構造

このようにして定義された XML 文書に対し、次のような問合せの例を考える。

Q1: つくば市の領域を表す多角形オブジェクトの情報を抽出せよ。

この問合せは、テキストと文書構造に関する条件のみを含んでいることから、構造化文書に対し提案された既存の検索モデルを用いて検索を行うことが可能である (記述例は後述)。一方、以下の問合せを考える。

Q2: つくば市の領域内の学校の名前をすべて列挙せよ。

この問合せは空間問合せの一種である範囲問合せを含んでいるため、XML 文書を単なる構造を持ったテキストと捉える従来の XML 問合せ言語では記述は困難である。そのため、Point, Polygon といった空間データを表す文書要素を空間データとして認識し、範囲問合せの条件を判定する述語を構造化文書の検索において統合利用できることが必要となる。

本研究の目的 以上の理由から、各種応用分野における XML データの効率的な検索処理のためには、ユーザ定義のデータ型や関数が文書検索機能に統合された形で利用可能であることが求められる。従来より、構造化文書に対する検索のために、構造化文書の構造と内容に基づく検索のモデルが多数提案されているが [6], ユーザ定義のデータ型や関数が含まれている場合については検討がなされていなかった。本研究では、ユーザ定義のデータ型、関数が含まれる場合において構造化文書を効率的に検索するためのアプローチを提案する。以下では、先に示したような地図データの XML による表現を想定し、空間的な条件も含めた形で文書検索を記述するための枠組みについて述べる。

2 構造化文書検索モデルへの拡張性の導入

構造化文書検索モデル [6] でサーベイされている構造化文書に対する検索モデルは、従来の文書検索のモデルの延長線上に位置し、文書の内容と構造に基づく検索条件を同時に指定して検索できる点を特徴とする。文書構造を考慮した文書要素の検索機能という点で、この種の検索モデルは、XML の領域における XPath [7] などの問合せ言語と対応づけることができる。このような検索モデルは、XML 文書に対する強力なデータ構造操作機能を有する XQuery [3] などの XML 問合せ言語に比べ表現能力の面で劣っているが、検索対象の文書要素の特定に機能を絞っていることから、一般により効率的な処理が可能であり、高レベルな XML 問合せ言語の処理系を実現する場合の検索サブシステムの実現においても利用できる。

Proximal Nodes モデル 本研究では、構造化文書に対する検索モデルに関する既存の提案の中から、表

An XML Retrieval Model Supporting User Defined Types
Yoshiharu Ishikawa[†], Takaaki Nagai^{††}, and
Hiroyuki Kitagawa[†]

[†] Inst. of Inf. Sci. and Elec., Univ. of Tsukuba

^{††} Graduate School in Sys. and Info. Eng., Univ. of Tsukuba

現能力と効率の面でバランスがよいとされる *proximal nodes* モデル [8] をもとに拡張を行うことで、ユーザ定義データ型と関数を含んだ構造化文書の検索を支援する。proximal nodes モデルの特徴は効率を意識した演算体系にあり、転置ファイルなどの索引に基づいて検索式の評価に必要な部分的な文書要素が特定できると、それをもとにボトムアップに検索結果が構成できる点にある。

proximal nodes モデルによる検索の例として、先に述べた問合せ Q1 に対する問合せ記述を示す。

```
Q1: Polygon in (
  (GeoFeature parent
    (category same "area")) is
  (GeoFeature with
    (name same "Tsukuba City")))
```

この問合せでは、category 要素の値が "area" であり、かつ、Property 要素の下に name 要素の中に "Tsukuba City" という文字の並びが含まれるような GeoFeature 要素を取り出し、それに含まれる Polygon 要素の検索を指定している。問合せの評価時には、GeoFeature などの文書要素や "area" などの文字列の出現は、それらに対応するリージョン（文書上のその文書要素の出現範囲）の集合に置き換えられる。parent などの演算子は、2つのリージョン集合を引数にとり、結果としてリージョンの集合を生成する。with, same, parent, is 演算子は、それぞれ、リージョン集合間の包含関係、一致関係、親子関係、積集合演算に対応している。

検索モデルの拡張：空間演算子の導入 先に示した問合せ Q2 を、proximal nodes モデルに対する拡張を導入して以下のように記述する。

```
Q2: name in (
  (GeoFeature with (type same "school")) is
  (GeoFeature with (Polygon spat::in Q1)))
```

Q1 の箇所には、先の問合せ Q1 の内容が展開される。spat::in の部分はユーザ定義の関数 (spat モジュールの in 関数) の呼出しを表しており、この関数の引数である Polygon と Q1 の評価結果のリージョンの集合は、それぞれ多角形の値の集合として解釈されて関数に与えられる。

proximal nodes モデルではテキストの領域間の包含関係などに基づいて演算子の処理体系を定義していたのに対し、上のような処理では空間的な領域間の包含関係などに基づく処理が必要となる。セマンティクスは異なっているが、両者とも包含関係に基づくことでは一致しているため、空間的な包含関係のユーザ定義関数の呼出し機構を適切に構築すれば、proximal nodes の評価機能をそのまま利用して検索処理が行えることになる。

検索モデルの拡張：概念階層の導入 proximal nodes モデルでは、文書構造の階層として複数の階層を併

用して検索処理を行うことを可能にしている。たとえば、章、節、段落といった論理的な文書構造の階層と、ページ、ページ内の位置といった物理的な階層を同時に利用できる。しかし、現実の問合せ処理においては、文書内に現れる階層構造だけでなく、現実世界に存在する階層的な情報も統合できると、その有用性が増大する。

ここでは、概念階層の例として

```
country -> state -> city -> district
```

という、行政単位の間階層関係を考える。このような階層関係に含まれるインスタンス間の関連は、外部のデータベースに含まれているものとする。このとき、「Q3: 春日という地区を含む市町村の名前を列挙せよ」という問合せは以下のように与えられる。

```
Q3: name child
  (city addr::parent
    (district addr::with "Kasuga"))
```

ここでは、addr モジュールにより、行政単位に関する概念階層の情報が検索可能であると考えている。このような機構を導入することで、たとえば「春日地区を含む市町村の領域内の小学校は?」といった、空間情報と概念情報を統合した検索も可能になる。

3 まとめ

本稿では、構造化文書に対する検索モデルの1つである proximal nodes に対し、ユーザ定義型とユーザ定義関数に関する拡張性を与えることで、空間データを含む XML 文書を処理するためのアイデアについて述べた。今後は、さらに本提案の詳細化と、空間索引などの利用による効率化について検討する。また、X²QL による G-XML 問合せ処理 [4] における問合せ効率化のためのサブシステム構築における本提案の利用についても検討したい。

謝辞

本研究は文部科学省科学研究費 (12480067 および 12780183) の助成によるものである。

参考文献

- [1] <http://www.w3c.org/XML/Schema>
- [2] <http://www.xml.gr.jp/relax/>
- [3] <http://www.w3c.org/TR/xquery/>
- [4] N. Shinagawa, T. Nagai, H. Kitagawa, and Y. Ishikawa. Querying Geographic Data in XML via eXtensible XML Query Language X²QL. in *Proc. Intl. Symp. on Asia GIS 2001*, Tokyo, Jun. 2001.
- [5] <http://gisclh.dpc.or.jp/gxml/>
- [6] R.A. Baeza-Yates and G. Navarro. Integrating Contents and Structure in Text Retrieval. *ACM SIGMOD Record*, 25(1):67-79, 1996.
- [7] <http://www.w3.org/TR/xpath>
- [8] G. Navarro and R. Baeza-Yates. Proximal Nodes: A Model to Query Document Databases by Content and Structure. *ACM TOIS*, 15(4):400-435, 1997.