データベースシステム 129－14
（２００３．１．２４）

XML ベースのコンテンツガイドシステム実現について

日高 東潮　　戸田 浩之　　小島 明　　片岡 良治　　星 隆司
日本電信電話株式会社 NTT サイバースペース研究所
{ hitaka.toshio, toda.hiroyuki, kojima.akira, kataoka.ryoji, hoshi.takashi} @lab.ntt.co.jp

概要
現在我々のグループでは XML に基づくメタデータ流通基盤のための研究開発を行っており、その中で近年ブロードバンドにおいて映像配信システムの普及と共に必要となるコンテンツガイドシステムの検討を行っている。XML で表現されたコンテンツ情報を登録管理し、高速な検索を実現するためには現在のところ RDB システムを使用するのが一般的である。しかし映像コンテンツを表現する XML の仕様は構造が複雑かつデータ量も従来の DB 管理データに比べ大きいという特徴があり、そのため RDB システムで効率良く扱うためには従来とは異なる DB アプリケーション開発のアプローチが必要となる。本稿では SQL-92 で想定される RDB の範囲で映像コンテンツの代表的なスキーマである MPEG-7[1]を例に、XML を扱う場合の問題を整理し、近年の DBMS 製品機能のうち、どのような機能がその問題を解消するのに有効であるかの検証を行った。

An aproach for construct the XML-base contents guide system.

Toshio Hitaka　Hiroyuki Toda　Akira Kojima　Ryoji Kataoka　Takashi Hoshi
NTT Cyber Space Laboratories
{ hitaka.toshio, toda.hiroyuki, kojima.akira, kataoka.ryoji, hoshi.takashi} @lab.ntt.co.jp

We've been researching on the platfome system for XML metadata circulation base, and our group is examining the contents guide system which is needed with the spread of broad band systems in recent years. Although it is common to use a RDB system now in order to manage the contents information data, it is hard to get quick response and storage flexibility by the reason of complicated structure on such kind of XML data. In this paper, we forcus on the problem in the case of treating XML for MPEG7 [1] which is a typical schema of contents, and in the range of DBMS assumed by SQL-92, we verify what function would be effective on solving the problem among DBMS product functions in recent years.

1. はじめに

近年ブロードバンドの普及と共に映像配信サービスが多く見られるようになり、複数のコンテンツ提供事業者からサービスプロバイダーがコンテンツ情報を納品しサービス提供するようなモデルを前提に、コンテンツガイドシステムが提案されてきている。それらの多くはレガシーな RDB による番組管理情報データ（コンテンツ情報）に基づいたものであるが、独自のフォーマットにより管理されたコンテンツ情報では情報を円滑に流通させる上でさまざまなフォーマット間の変換ルールをすべて理解する必要があり、現実的ではない。そのための共通化フォーマットとして、MPEG7 や TV-Anytime 等のメタデータ仕様が提唱されているが、これらのデータは XML としては非常にサイズが大きく、タグのネストが深い複雑な構造であり、市販されている RDB の上に単純にデータを格納し、高速な検索アクセスを実現するのは困難である。

本稿では、実際にコンテンツガイドシステムを構築

する場合に必要となる機能を 2 章で整理し、その上で市販の RDB を対象に高速な検索性能を提供するための考え方について 3 章で整理する。その上で、現在 XML データ流通プラットフォームの一応用例として研究開発中のコンテンツガイドシステムの上に実装した場合の性能等について 4 章にて言及する。

1.1. 関連技術について

XML データの管理用 DBMS としては、近年 XML ネイティブ DBMS などが実用化されつつあるが、現在のところ性能面や製品の熟成度の観点、さらには共通的なスキーマに基づいた多数のデータを管理する状況が多いことなどから RDBMS が使用されることが多い。XML を格納する RDB 設計については従来幾つかのアプローチが提案されており、それらはスキーマ依存･非依存の 2 種類に大きく分類される。ここではそれぞれについて説明する。

まずスキーマに非依存のものとしては文献[2]で紹介されている Edge 分解方式のものがあり、これは XML タグのネスト関係をすべてテーブル間の外部参照キーで結んだ多数のテーブル分解で構成する方法である。本方式は構造が複雑な XML ではテーブル数が増加し、検索の際に複雑な JOIN 操作が必要となることによる処理コスト増大の傾向がある。スキーマの依存性に関わらず、多くの手法で本方式の考え方が基本とされている。

スキーマ依存のものとしては文献[3]の shared mapping 方式のものが代表としてあり、これは Edge 分解方式での検索コスト増大を避けるためスキーマの情報を積極的に利用し、JOIN 頻度の高い個所についてはデータの二重化を許容し結合表を予め作成しておく方法である。これは検索コストを下げる一方でスキーマに対する依存性が高いことからスキーマ変更に対し柔軟性が低いという傾向を持つ。

その他にも DBMS のオプティマイザ出力を利用して複数の XML-RDB スキーママップを用意し、実際のアクセスによるそれぞれのマップ使用に対する検索コストを参照してチューニングし、最適な shared map を生成する文献[4]のような方法もある。

以上の手法は XML に対する汎用的な格納･検索方式を提供するものであり、本稿ではこれらの方式のうち有効なものを利用してコンテンツガイドで求められる機能に特化したシステムを検討する。なお本稿では対象とする XML スキーマは固定されている問題を取り扱うため、文献[4]で触れられている自動チューニングについては言及しないものとする。

1.2. 本文の構成

本文はまずコンテンツガイドシステムにおける検索要求と、それが対象とするデータモデルについて整理し、それを SQL-92 ベースのレガシーな RDB ベースでの実装検討について述べる。

次にその検討で検出した問題点について、近年の DBMS 製品で有効と思われる実装機能について述べ、それを利用することの優位点と、問題点について述べる。

2. 課題とアプローチ

2.1. 想定するコンテンツガイド機能とデータスキーマについて

本稿で想定するコンテンツガイドシステムに求められる機能を整理するにあたり、現在のコンテンツ配信システムなどで提供されているコンテンツガイド機能[5][6][7]を整理したところ、現時点では以下のような機能でほぼ網羅される。

- ジャンルによるコンテンツ分類
- タイトル名などによるキーワード検索（部分一致）
- 項目を指定しないフリーキーワード検索
- 公開日時などの数値情報に対する範囲検索

これらは一般的に固定スキーマによる RDB 管理されたコンテンツデータに基づく検索サービスを提供するものであるが、近年映像データの作成過程、もしくは映像そのものからの特徴量抽出技術の向上[8]により、セマンティックなものから数値情報までさまざまなデータがコンテンツデータとして提供可能な環境の下地ができつつある。それらは将来的に XML のような半構造スキーマにより1元的に管理されることが考えられるものであり[9]、その管理用スキーマの1つとしてとして現在 MPEG-7 などが議論／規格化されている。

前述のサービス機能分類もとに整理すると、MPEG-7 をベースとしたのコンテンツガイドシステムで要求される XML への検索機能としては下記のように分類される[10]。

① コンテンツ XML の属性を指定した部分一致キーワード検索（ジャンル検索はジャンル名に対する検索として整理する。）

② コンテンツ XML の属性を指定した数値に対する範囲検索（映像特徴量などの属性に

対する検索もこの数値範囲検索として整理する。）

また MPEG-7 はデータ投入の汎用性を考慮した規格であるため、非常に膨大なサイズのスキーマにて規定されている。本稿ではコンテンツガイドとしての利用を想定したインスタンス仕様として、別紙に示すスキーマツリー図を想定する。

2.2. RDB 実装方法例と問題点について

本節では、まず MPEG-7 データを SQL-92 で想定されているレガシーな RDB に実装するケースについて検討する。

MPEG-7 データは一般的に RDB にて管理されるデータよりも複雑な構造になりやすい点、並びに DB 格納データとしてはかなり大きなデータ（1 データあたり、約 10KB 以上。RDB のカラムはもちろん、レコードとして比較しても大きい部類。）である点を特徴とする。そのため、MPEG-7 を RDB 上に格納するため DB スキーマにマッピングする場合、工夫が必要となる。考えられるマッピング方法としては基本的に下記の 2 種類となる。

① MPEG-7 データを RDB 上の可変長 1 カラムに対して格納する方法

② MPEG-7 データをタグごとに分解し RDB 上に複数の Relation Table を作り、対応付けして格納する方法（Edge 分解、 Shared Mapping）

それぞれについての格納方法と検索方法、並びにメリット・デメリットについて整理する。

① MPEG7 データを RDB 上の可変長 1 カラムとして格納する場合

［格納方法］

RDB 上に番組 ID（ProgramID、図中 PID と略記）と MPEG-7 文章全体の 2 つのカラムから構成されるテーブルを用意して格納する。この MPEG-7 文章全体を格納するカラムは、1 つのデータサイズが通常 10KB 程度以上になるため、RDB でよく使用される可変長文字列型での格納がデータ長制限により使用できない事が多いため、Large Object 型の特殊なデータ型を定義して格納することになる。（図1）

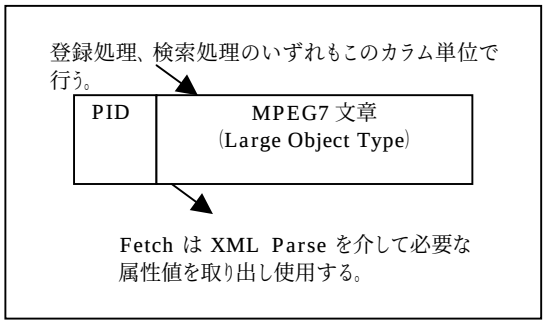


図1：方式1のデータ格納方法の概略図

[検索・更新方法]

文字列の部分一致条件による検索を行う。

更新の際には、検索と同様の方法で更新対象データを確定した上で、XML データ全体を破棄して新しい XML データに入れ替えることで実現する。これは、SQL による更新単位がカラム単位であることに基づく制約であるあ。

［メリット］

a. 登録の際に特に XML の Parse などを考慮せずにデータ挿入が可能。

b. XML の内部構造に依存しないので、汎用性が高い。またシステムの運用開始後にスキーマの変更を行うことも可能。

［デメリット］

c. 検索において、「あるタグ指定の個所の値が X である」という条件を想定した場合、事前にタグごとに分解してインデックスを作らない限り、全件検索を行わなければならず、検索コストが N オーダー必要となってしまう。

d. 検索での fetch 対象として「あるタグ指定の個所の値」が指定された場合、一旦メモリー上に格納されている MPEG-7 文章をすべて展開し、さらに Parse 処理を行いデータ抽出をしなければならない。一般的に XML 文章の Parse 処理には 1 件あたり秒オーダー近い処理時間が必要となるため応答性能を出すことは困難である。

e. データの更新を行う場合、データ全体を更新する必要がある。その場合、データの書き換わる範囲は全体のほんの一部にも関わらず、更新時に発生する DBMS 内のジャーナル情報はデータ全体の情報が書き込まれるため更新の効率が非常に悪

い。更新条件として「あるタグ指定の個所の値を X という値で更新」という形で与えられた場合、データ全体の値を更新する関係上、一旦メモリ上にデータ全体の値を展開して必要個所の書き換えを行い、そのデータを DB 上へ書き込むという操作が必要となり、指定された更新条件に比べ、更新量が極めて大きくなる。

② MPEG-7 データをタグごとに分解して RDB 上に複数の Relation Table を作り、対応付けして格納する方法

［格納方法］
MPEG-7 データをタグごとに分解し、ProgramID を外部キーとする複数の RDB テーブル（以下、アトリビュートテーブル）にマッピングする。この場合、カラムとしては方式①に比べ非常に小さくなるため、通常の可変長文字列型として格納することができる。また、数値情報などについては、数値型として格納する。（図2）

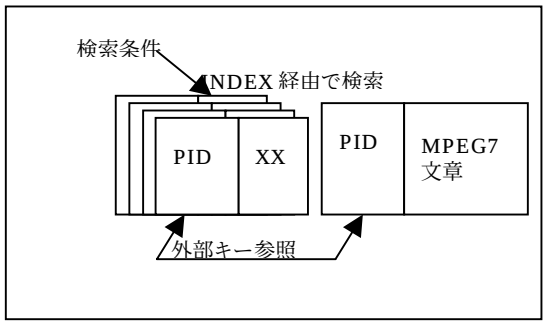


図2：方式2のデータ格納方法の概略図

[検索・更新方法]
検索は検索条件の対象となるアトリビュートテーブルに対して行い、ProgramID が確定した段階で返却に必要な属性すべてに対し、ProgramID を外部参照キーとした多段 JOIN によって結合させ、返却する。
更新の際には更新が必要な個所だけの UPDATE を行う以外に、更新が終了するまでの間、該当 ProgramID を外部キーとして持つレコードすべてに更新ロックをかけるなどの操作が必要となる。

［メリット］
a. 検索の際に「あるタグ指定の個所」が指定された場合、必要最低限の個所をディスクからメモリへデータ転送するだけで処理が実現できるため、ディスク I/O の量を抑えることができる。
b. 検索の際に「あるタグ指定の個所の値が X である」という完全一致、もしくは前方一致条件指定を与えられた場合、タグごとに分解された値に対してインデックスを張ることで、高速な検索性能を得る事が可能となる。
c. データの更新を行う場合、更新条件として与えられた個所のみを更新すれば良いため、更新処理のうち特にディスク I/O の量を抑えることか可能となる。

［デメリット］
d. 検索の際には複数のテーブル間で多段 JOIN 操作が発生するため、MPEG-7 などのケースでは最悪 JOIN 段数が 10 段にも及ぶようなケースが発生し得るため、RDBMS に対して高負荷を招き応答性が著しく悪化することがあり得る。このことによる性能劣化は容易にメリット a を上回る事が多い。前節で述べた Shared Mapping 方式を導入することである程度の回避は可能ではあるが、スキーマ変更に対する柔軟性、要求される検索パターンに対する柔軟性を阻害するという問題がある。
e. 検索の際の条件指定が完全一致、もしくは前方一致であれば RDB インデックスをうまく組み合わせることで高速な検索が可能であるが、Web サービスなどの場合、かなり多くのケースで部分一致検索に対する範囲検索が要求されることが多く、その場合は RDB インデックスが使用できず、全文検索のロジックが選択されてしまい検索コストがかかるケースがある。
f. 新規にデータ登録などを行う場合、XML データの Parse 処理が入るための処理時間が必要になる。
g. データ登録の際にはテーブル間に跨った複数テーブルでの同時 commit が必要となるため、更新ロック数が多くなり、DBMS のロック制御に対して負荷をかける。

以上 2 つの方法を考えた場合、どちらも実際のシステム構築においては検索性能の観点だけでもさまざまな問題点を持つ。また、方法1と方法2を組み合わ

せた実装方法も考えることができるが、これはアプリケーションの参照方法に依存して組み合わせを考えねばならず、アプリケーション毎の設計が必要であるため、現実的ではない。

本稿では、コンテンツガイドシステムという性格上、データの登録処理の性能より、検索処理の性能を優先するという指針を加え、上記問題点の解消方法を検討する。

3. 提案方式

前節で述べた通り、レガシーな RDB 機能の範囲では MPEG-7 データを扱うためには幾つかの問題がある。そこで、近年製品として用意されている機能から、標準的に搭載されつつある全文検索インデックスを拡張した XML の構造に特化したインデックス機能を使用することを主軸に考え、不足する機能については順次述べることにする。

3.1. 全文検索インデックスの使用

近年、RDBMS 製品の多くが B-Tree インデックスや Hash インデックスに加え、全文検索用のインデックスを標準的に提供しつつある。さらに全文検索インデックスも XML に対する機能拡張が加えられ、XML のタグ付き文書に対応した検索機能をも提供する方向にある。今回、この全文検索インデックスを 2 章で述べた方法①に対して加えることを考える。

これにより方式②で対応できなかった文字列に対する部分一致検索の高速実行を可能とする。(図3)

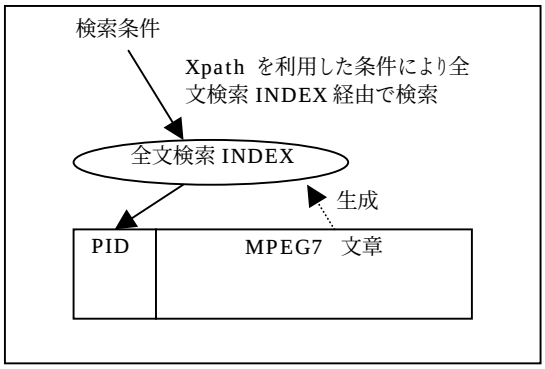


図3:提案方法の基本となるデータ格納概略図

3.2. 全文検索インデックスでは対応できない検索のためのアトリビュートテーブルの利用

前節で述べた全文検索インデックスの使用では、実際にはアプリケーションで必要とする検索機能を提供できない個所がある。それぞれについて説明の上、それらに対する対処方法について検索する。

3.2.1. 数値の範囲指定検索

一般的に全文検索インデックスでは数値の範囲指定検索には対応しないが、コンテンツガイドにおいてはセマンティックな“公開年月日”などに対する時間の範囲指定、画像特徴量に対する範囲指定などまで幅広い数値の範囲指定機能の用途がある。

そのため、本稿ではこれに対応するため 2 章方式②で述べたアトリビュートテーブルを数値の個所に作成して対応する。このアトリビュートテーブルには RDB インデックスを作成することで、検索条件に該当する MPEG-7 文章の確定までを高速に行うことができる。(図4)

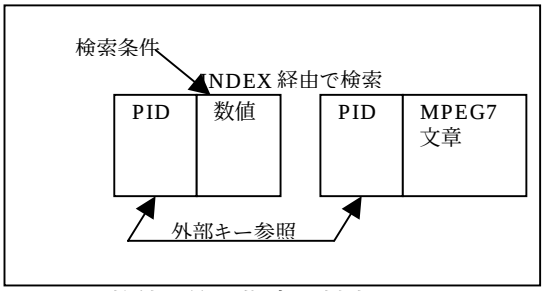


図4:数値の範囲指定に対応するための
アトリビュートテーブルの概略図

3.2.2. XML 上分岐条件検索

映像コンテンツを対象とした検索では、よく発生する検索としては「監督が“○○○”であるような映像」という検索条件が考えられる。この場合、XML としては分岐条件になり、XPath で表記すると、

＜//Creator[Role/@href=
”監督”]/Agent/Name/GivenName/text()=”○○○”＞

という検索条件になるが、現状市販されている製品の多くがこの検索条件に対応しておらず、そのためこのような条件検索では全文検索インデックスで対応することは難しい。そのため、本稿では前出の Shared Mapping 方式によるアトリビュートテーブルを作成し対応する。(図5)

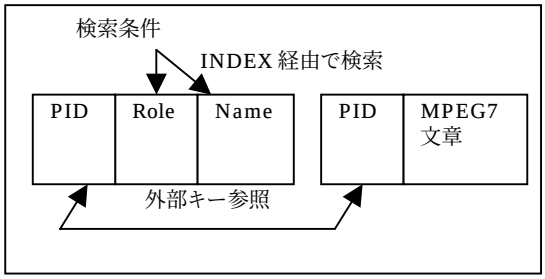


図5：Xpath 分岐条件に対応するための
アトリビュートテーブルの概略図

3.2.3. 検索フィルターの導入

これまで述べた方法では、アプリケーションから見た場合に検索条件によって、全文検索を使用する個所とアトリビュートテーブルの RDB インデックスを利用する個所とが分かれており、下位の実装を理解しなければアプリケーションが作成できないという問題がある。また XML はスキーマ定義が柔軟である反面、システムの運用後にスキーマを拡張するなどの要求も発生しやすい。

そのため、本稿ではアプリケーションから要求される検索条件と内部実装の検索条件を自動変換するための検索フィルターを用意して対処している。

このことにより、新たなアプリケーションの追加により必要な検索のパスが新たに発生した場合、このフィルターに新しい検索条件の対応ルールを追記・変更することで、柔軟に対応することを可能とする。（図6）

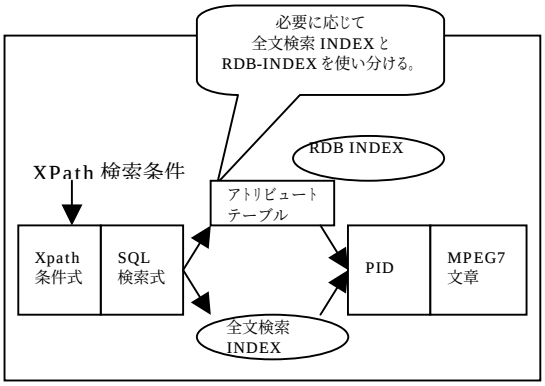


図6：検索フィルターの概略図

4. 実装と評価

4.1. 実装について

今回の提案手法の実現性、有効性を確認するためプロトタイプを作成した。システムの概要を図7に示す。Web サーバー上に MPEG-7 の格納・管理・検索機能を提供するプラットフォーム部を置き、XML をインタフェースとする Web アプリケーションに対して検索機能を提供するという構成である。

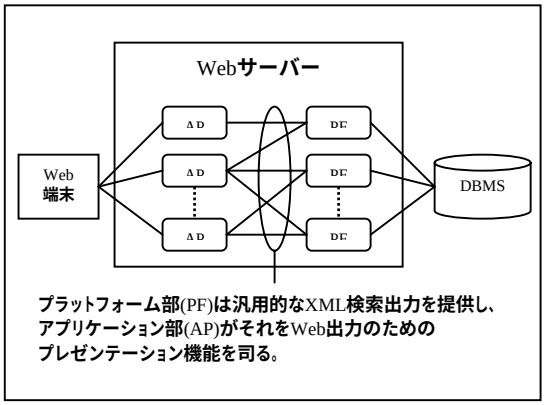


図 7：実装システムの概要

高速応答性のために、今回は AP へ返却する XML データは外部要求に応じてその都度 MPEG-7 を Parse して要求に応じた形態に再構成することはせず、登録の際に AP で必要とする形態の XML データを ProgramID を外部キーとした MPEG-7 サブセットを用意して対応している。

アプリケーションとしては、番組、シーン単位で持つ各種属性に対するキーワード検索機能、数値の範囲検索機能、その他画像特徴量に基づく類似度検索機能などを提供する。

4.2. システムの評価

4.2.1. 性能評価

本稿で提案したシステムの有効性を検討するため、3 章で提案した方法に基づくコンテンツガイドシステムを試作し、従来から開発研究を行っていた 2 章の方式②に順ずる映像コンテンツガイドシステム[7]との間で比較評価した。

監督名を指定しての検索を方式2のものと本稿で提案のものに対し同様に行い Web 端末での応答速度を確認したところ、方式②では 10 秒近い検索時間を必要としたのに対し、本提案方式では 2 秒以下

での検索処理を実現した。

これについては詳細に解析を行ったところ、DBMS に対して方式②は多段数の JOIN を行っていることに加え、文字列の部分一致検索が要求されたためにアクセスパスとしてインデックスが使われておらず、全件検索のロジックが使われていた。そのため DBMS 上で JOIN をする際に一般的に使用される Join で大量のデータソートとマージ処理が起こり、そのことから上記のような性能の差を招いたものと推察される。

4.2.2. 機能面の評価

本稿で提案した方式では、
① キーワードの部分一致検索での高速応答性
② 数値の範囲指定に対する検索機能
の2点について実現を可能とした。
また、更新処理の面では、
① スキーマ拡張に対する柔軟性の実現
② 更新の際のロック数の抑止
などについては解消できているが、更新量を必要最小限に抑えるなどの対応については言及できていない。ただし現在考慮しているコンテンツガイドシステムについては1データあたり10KB程度のデータ量であり、更新頻度も低いことから、現時点で想定している使用方法であれば問題にならないと考える。

5. まとめ

MPEG-7 に代表されるような複雑な構造の XML データについて、SQL-92 の機能を前提に問題点を明らかにし、その上で市販製品のサポートする機能を有効に利用する方法について提示した。

今回提示したのは MPEG-7 データを格納し、コンテンツガイドシステムで想定される検索条件に対し高速に該データを確定するまでの処理について言及したが、実際にはデータが確定した後、XML から必要な個所を Parse して出力する処理が必要となる。現時点では Parse を行うと Java 等で実行した場合秒オーダー近い処理時間が必要となるため、現時点では必要となる出力形態をデータの登録時に作って保持しておくなど、多重のデータ保持を許容した対処が必要となる。この点については今後XML の Parse 処理の高速化を含め整理が必要である。

また、本稿では検索時に特化して言及したが、MPEG-7 データの更新については、機能としては柔軟な方法として与えた一方で、更新の際にジャーナルを大量に生成してしまい、性能面で問題を残している。これらの点については今度、継続的に検討を行う予定である。

[参考文献]

[1]"Text of ISO/IEC FDIS 15938-5 FDIS (MPEG-7 MDS)" , 2001.

[2]D.Florescu and D.Kossmann, “Storing and querying xml data using a rdbms,” IEEE Data Engineering Bulletin, Vol.22, no.3, pp.27-34,1999.

[3]J. Shanmugasundaram, K.Tufte, C.Zhang, G.He, D.J.DeWitt, and J.F.Naughton, “Relational databases for querying xml documents: Limitations and opportunities,” in VLDB’1999, Edinburgh, Scotland, UK(M.P.Atkinson, M.E.Orlowska, P.Valduriez, S.B.Zdonik, and M.L.Broodie,eds.), pp.302-314, Morgan Kaufmass,1999.

[4]P.Bohannon, J.Freire, P.Roy, and J.Simeon, “From XML schema to relations:A cost-based approach to XML storage,” in ICDE, 2002.

[5] “BROBA”, http://www.broba.cc/

[6] “111.tv”, http://www.111.tv/

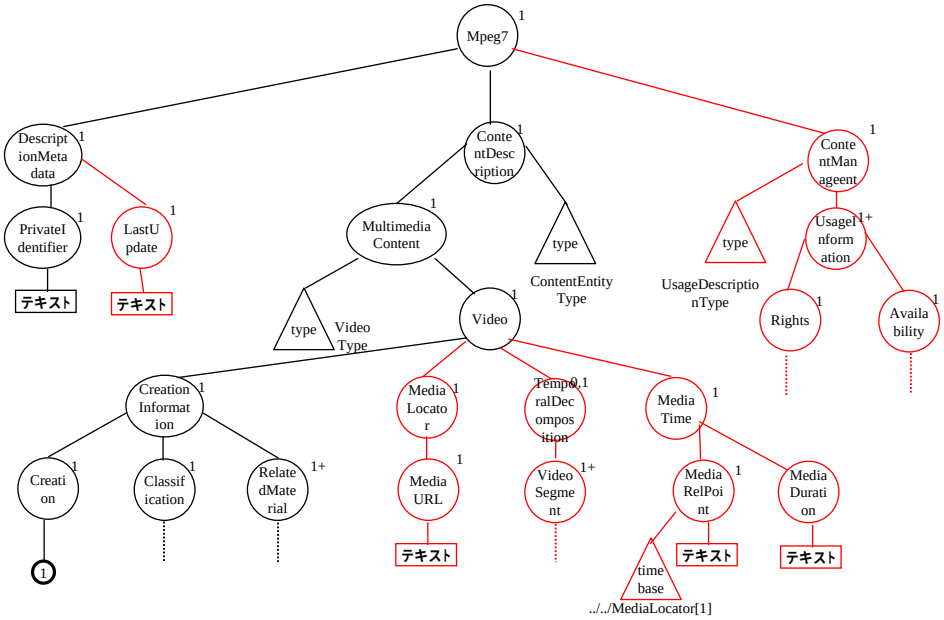
[7] “@フューチャー”, http://www.e-movie.co.jp/

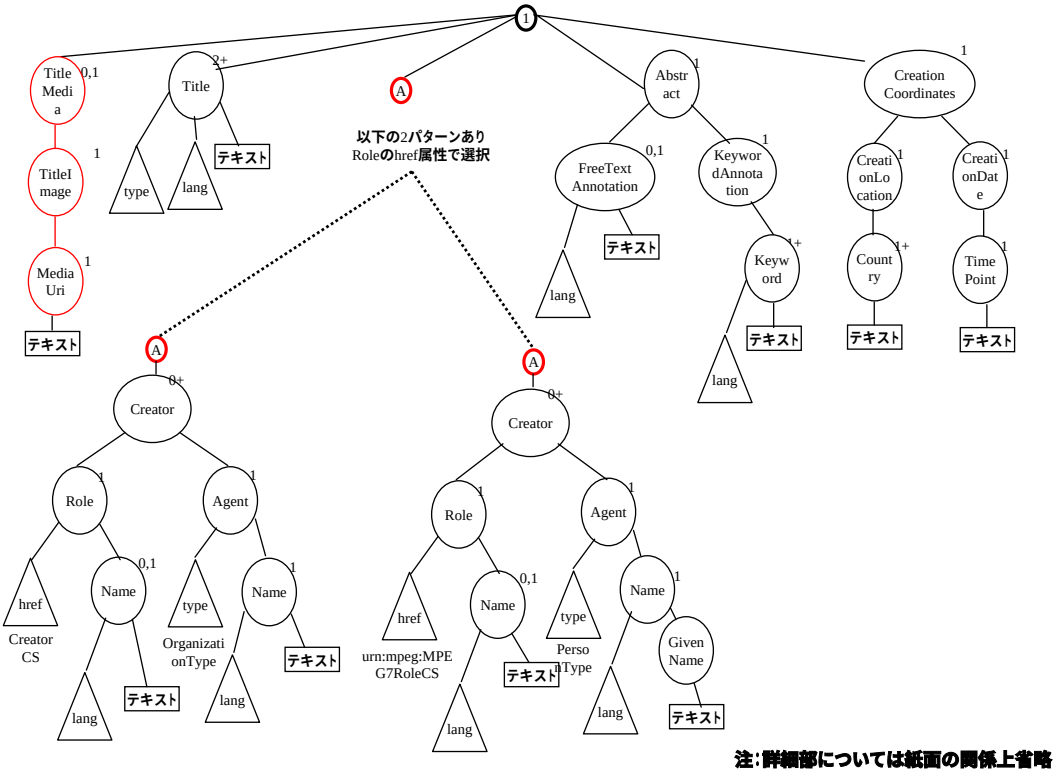
[8]谷口行信ら，"SceneCabinet:映像解析技術を統合した映像インデクシングシステム"，信学論(D-Ⅱ), Vol.J84-DⅡ, No.6, pp.1112-1121, 2001.

[9]"Text of ISO/IEC FDIS 15938-3 FDIS (MPEG-7 Visual)" , 2001.

[10]戸田浩之ら，"映像配信サービスにおける状況適応型検索システムの提案"，情報処理学会研究報告, DBS-127, pp.121-128 , 2002.

別紙


Mpeg7
DescriptionMetadata
PrivateIdentifier
LastUpdate
テキスト
テキスト
ContentDescription
Multimedia Content
type
Video Type
Video
type
ContentEntity Type
CreationInformation
Creation
Classification
RelatedMaterial
MediaLocator
MediaURL
テキスト
TemporalDecomposition
VideoSegment
MediaTime
MediaRelPoint
time base
../../MediaLocator[1]
テキスト
MediaDuration
テキスト
ContentManageent
type
UsageDescriptionType
UsageInformation
Rights
Availability


TitleMedia
TitleImage
MediaUri
テキスト
Title
type
lang
テキスト
以下の2パターンあり
Roleのhref属性で選択
Abstract
FreeTextAnnotation
lang
テキスト
KeywordAnnotation
Keyword
lang
テキスト
CreationCoordinates
CreationLocation
Country
テキスト
CreationDate
TimePoint
テキスト
Creator
Role
href
CreatorCS
Name
lang
テキスト
Agent
type
OrganizationType
Name
lang
テキスト
urn:mpeg:MPEG7RoleCS
PersonType
GivenName
注：詳細部については紙面の関係上省略