ナレッジマネジメント向けXML処理エンジン

XML Processing Engine for Knowledge Management

服部　雅一　HATTORI Masakazu
末田　直道　SUEDA Naomichi

組織内に存在する情報群をXML（eXtensible Markup Language）ベースで管理・蓄積し，オンラインで要求したXMLデータを高速に検索・加工できるXML処理エンジンを開発した。このエンジンには，個人の分析プロセスをデータベース（DB）化し，組織内で共有，再利用できる機能も付加されている。このエンジンを用いることで，クレーム情報分析，日報や特許など各種の共有分析システムを効率的に開発することができる。

Toshiba has developed an XML processing engine which manages and accumulates XML-based information in an organization. This engine can search and process XML data on-line at high speed. Individual analysis processes are automatically stored in this engine. These processes can be shared and reused by other users in an organization. Using this engine, various types of information-sharing and analysis systems, such as claim information, daily reports, and patents, can be efficiently developed.

1 まえがき

現在，IT（情報技術）の進化によって，ばくだいな量の情報が容易に入手できるようになった。しかし，情報が大量に存在していても，それをうまく活用できなければ意味がない。ナレッジマネジメントが提唱されるのも，このような背景が存在するからである。

次世代のナレッジマネジメントの中核技術として期待される技術がXMLである。XMLは，拡張性と連携性を備えた標準の情報記述言語である。従来の文書検索において，「検索が効率的でない」，「膨大な情報群から新たな情報の抽出が困難である」など，ユーザーが感じている問題にXMLは対応することができる。

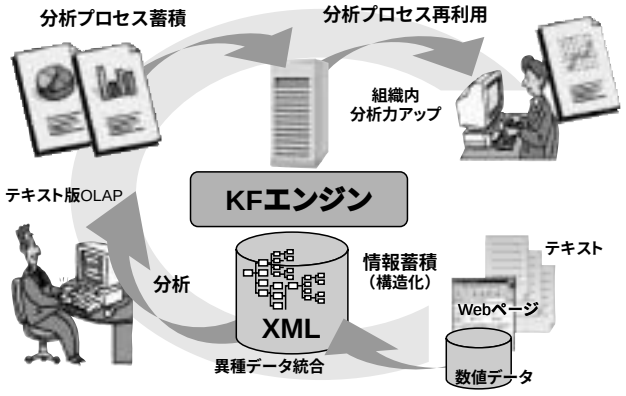

図1. XML処理エンジンに基づくナレッジマネジメントシステムのコンセプト　情報の蓄積，分析，分析プロセスの再利用までサポートする。

Concept of knowledge management system using XML processing engine

もちろん，XMLは単なる情報記述言語なので魔法の力を持っているわけではない。われわれは，XMLの効果を最大限に引き出すエンジンを開発している。それがKF（Knowledge Factory）である。これは，XMLをXML DBとして構造的に蓄積し，XML DB上で高度に検索・加工できるエンジンである（図1）。

2 ナレッジマネジメントへのXML適用

2.1 XML

現在，WWW（World Wide Web）上で情報を配布，閲覧するための言語としてHTML（HyperText Markup Language）が普及している。しかし，HTMLの問題点として，以下のことが挙げられる。

(1) タグは，情報のレイアウトを表現する機能として固定に定められている（例えば，<p>は改行を意味する）。

(2) 情報の構造とスタイルが一体化しているため，プログラム処理，再利用が困難である。

このような限界を解決，緩和するために登場した言語が“拡張可能なマークアップ付け言語”のXMLである。これは，標準化団体W3C（World Wide Web Consortium）によって仕様が制定されている。XMLの特長を以下に挙げる。

(1) ユーザーが用途に応じたタグ定義が可能である。文書に構造情報を付加することが可能で，プログラム処理，再利用を容易とする。

(2) タグの種類及びその構造，DTD（Document Type Definition）やXMLスキーマにより定義可能であるが，省略してもXMLデータとして成立する。

例えば，特許情報をXMLで表現したものを図2に示す。

```
<特許リスト>
<特許>
  <タイトル>情報検索装置</タイトル>
  <出願者>東芝</出願者>
  <出願番号>特願平10-××××</出願番号>
  <出願日><年>1998</年><月>×</月><日>××</日></出願日>
  <要約>
    情報の提示形式の変更が利用者側の観点で自由に行え、
    情報活用の範囲が広がるとともに、情報活用の促進が図れる
    <キーワード>DB</キーワード>
    を提供する。
  </要約>
   <キーワード>XML</キーワード>
   <キーワード>検索</キーワード>
</特許>
</特許リスト>
```

図2．XMLデータの例　特許リストをXMLで表現したものである。
Example of XML data

これをブラウザに表示させたい場合は，スタイルシートを用いてHTML形式に変換することもできる。

2.2　XML適用のメリット

現在，XMLは，企業間でWWWを経由してやり取りされる情報の記述言語として，特にEC（Electronic Commerce：電子商取引）分野で脚光を集めている。一方，社内情報の共有，再利用するための仕組みとしてXMLを利用する動きも広まりつつある。XMLには，以下の問題への対応が期待されている。

(1)　検索が効率的でない　　従来，文書検索といえばフルテキストサーチに代表されるキーワード検索であった。この手段は，文書とキーワードの比較で検索を行うため，精度が十分に得られない問題が発生した。

(2)　膨大な情報群から新たな情報の抽出が困難　　従来，関連する情報群を分類整理したり，情報間の関係を組織化したりして新たな情報を抽出することは困難であった。

(3)　定性データと定量データの統合が困難　　営業部門であれば，売上げデータ，営業活動情報，顧客情報など多種多様な定量／定性データにアクセスし，検索・加工を行う必要がある。しかし，リレーショナルデータと文書データでは，インタフェースが大きく異なり，統合が困難であった。

われわれは，XMLとXMLの効果を最大限に引き出すエンジンKFを用いることで，以下の効果を期待できる。

(1)　高精度な検索が可能　　論文を筆者タグやタイトルタグなどから構成されるXMLデータで記述すれば，筆者／タイトル／日付などから各条件を絞り込んで検索精度を高められる。また，検索結果についても必要な項目（筆者とタイトルなど）を一覧表示することも可能になる。

(2)　膨大な情報群の整理が可能　　XML文書と概念情報を組み合わせて分類して，複数文書を多次元テーブル上に表示したり，特定条件の文書発生件数を時間軸とともにトレンドグラフ表示したりすることが容易になる。このような操作をオンラインで実行することで，新たな関係を抽出することができる。

また，自然言語処理技術と組み合わせれば，分類項目を自動的に生成／追加することも可能になり，収集した文書を様々な角度からアクセスできる。また，動的に行うことで最新のデータを活用することができる。

(3)　定性データと定量データの統合が容易　　XMLは記述力の高い言語であり，リレーショナルデータや文書データを変換してXMLで統一することが可能である。KFを利用すれば，同一のインタフェースで定量／定性データをまたがった検索・加工が容易になる。

(4)　高度なナレッジマネジメントアプリケーションソフトウェア（以下，アプリケーションと略記）開発の効率化が可能　　4章で示すようなアプリケーションにおいて，処理の多くはKFの機能（コマンド）を用いて実現できる。われわれが構築した試作システムにおいて，開発コストは数分の一に低減された。

3　エンジンの概要

3.1　全体構成

KFは図3のとおりに実装されている。現状の実行環境は，

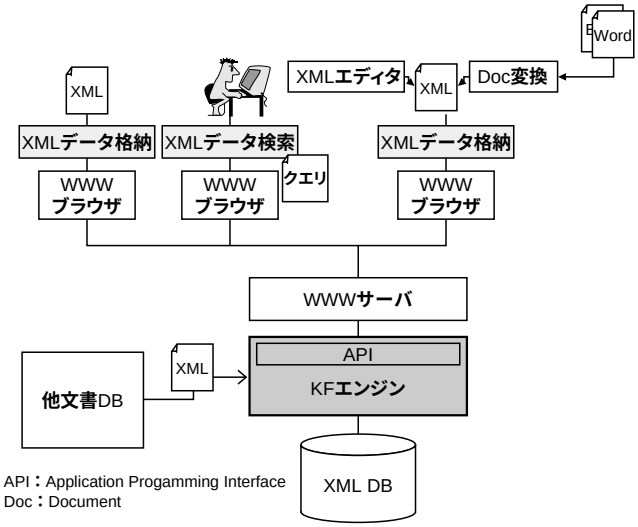

図3．全体構成　ユーザーは，WWWを通してXMLデータの格納・検索を行う。他の文書DBからの投入も可能である。
Software configuration

WWWサーバのバックエンドで動作している。

3.2 機能概要

KFの機能は格納機能と検索機能とに大きく分類される。

(1) 格納機能　ユーザーやアプリケーションから作成されたXMLデータを，XML DBへ格納する機能である。特長は次のとおりである。

(a) KFにおけるDBの論理イメージは，大きな階層構造を持ったXMLデータである。**図4**は，特許データや週報データなどが集まったXML DBのツリー表現の例を示している。部分的なXMLデータをアクセスするには，パスというアクセス手段を提供する。このような特徴により，幅広くXMLデータを検索・加工することが可能になる。

(b) XML DB上の一部にXMLスキーマを設定することで，妥当性のチェックが自動的に行える。現在は，XDR(XML-Data Reduced)というスキーマ言語をベースにしている。

(2) 検索機能　XML DBから検索条件を指定して条件に合致するXMLデータを検索する単純機能と，XML DBから加工条件を指定してXMLデータ群から新たな複合コンテンツとしてXMLデータを作成する複合検索機能に分類される。検索条件は，KF-QL(KF Query Language)という検索言語で記述される。特長は次のとおりである。

(a) 高度な加工に優れたKF-QLを実装し，複合コンテンツをダイナミックに作成することが容易である。XMLデータ群から集計や分類などを行う場合，特に能力が発揮される。**図5**は，「XML文書と概念情報を組み合わせて分類して，複数文書を多次元テーブル化する」クエリ(Query)の例である。

(b) KF-QLはXMLデータで表現されており，KF-QLで記述されたクエリ自体をログとしてXML DBに格納することも可能である。過去のクエリを蓄積，再利用するようなアプリケーションを容易に構築することができる。

(c) KF-QLには，HTMLにおけるリンクと同じように外部クエリへのリンク機能がある。XML DBを分析したい要求があるとき，XMLデータの一連の加工プロセスをクエリで実装しておけば，XML DB上で多次元分析ツールが構築できる。

3.3 実装技術

KFの実装技術として特徴的な2点を挙げる。

(1) ネイティブXMLデータ格納技術　一般にXMLデータを格納するには，RDB(Relational DataBase)やOODB(Object Oriented DataBase)など，従来DBMS(DataBase Management System)が使われるのが普通である。KFでは，XMLデータが持つ階層構造をファイルイメージにダイレクトにマッピングした方式により，多種多様なXMLデータを特別な変換なしに格納することができる。

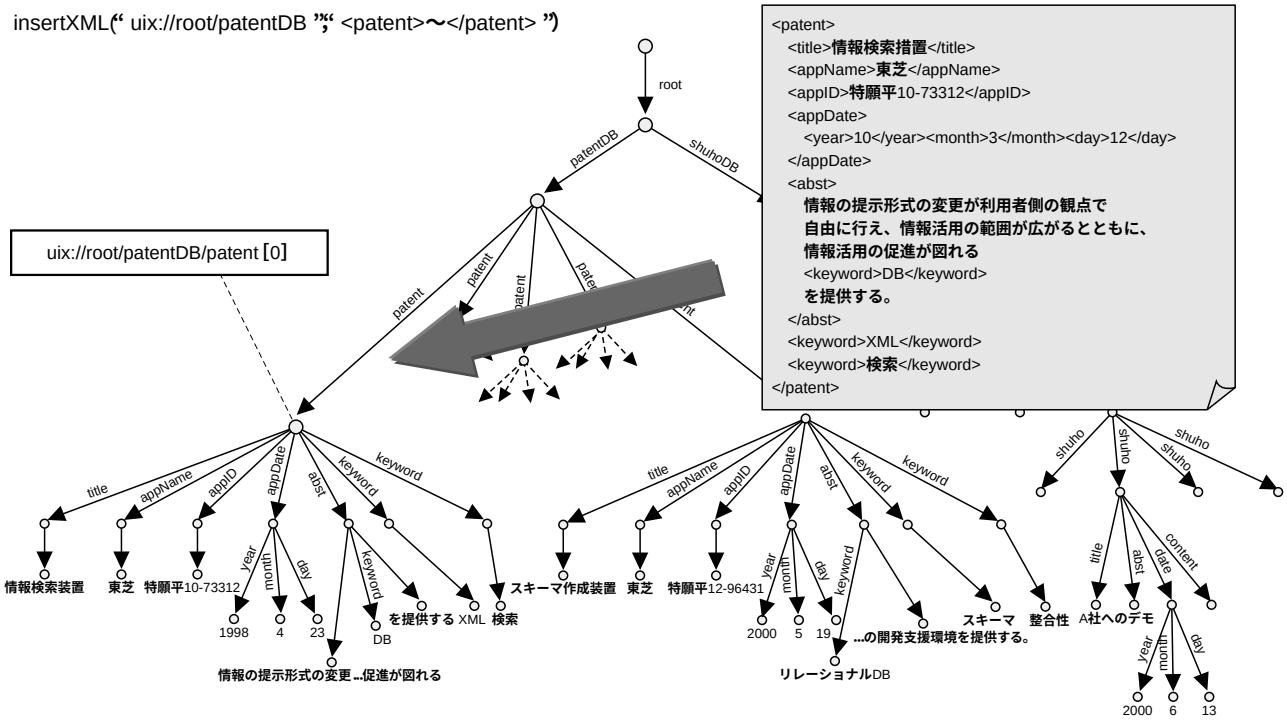

図4. XML DB**の論理イメージ**　insertXMLという挿入コマンドを使って，指定パス以下にXMLデータを格納する。
Logical image of XML database

```
<kf:query>
  <kf:select>
    <result>
      <タイトル>$t</タイトル>
    </result>
  </kf:select>
  <kf:from path=" http://www.a.b.c/PatentDB.xml ">
    <特許DB>
      <特許>
        <タイトル>$t</タイトル>
        <キーワード>$k</キーワード>
      </特許>
    </特許DB>
  </kf:from>
  <kf:from path=" uix://root/概念辞書 ">
    <概念 name=" 周辺装置 ">
      <概念 name=" $x "/>
    </概念>
  </kf:from>
  <kf:where>
    <kf:or>
      <kf:cmp op=" EQ "param1=" $k "param2=" 周辺装置 "/>
      <kf:com op=" EQ "param1=" $k "param2=" $x "/>
    </kf:or>
  </kf:where>
</kf:query>
```

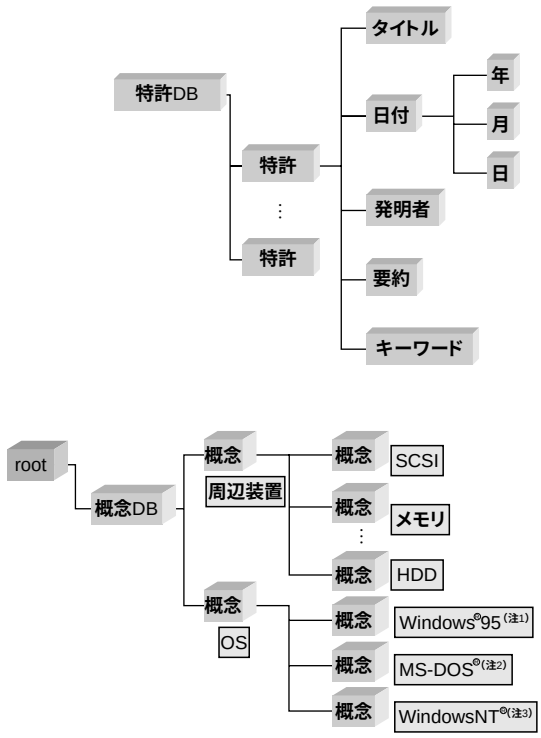

図5．クエリデータの例　概念情報を使って特許データを分類するクエリの例を示す。
Example of query data

⑵　**検索最適化技術**　　高度な加工に優れた検索言語ゆえに，ファイルアクセスを含めた処理時間の増大が問題となる。そのため，検索処理において，検索要求を内部グラフに変換したうえで最適な実行プランを生成する手法を開発した。

XMLデータ格納時，語彙（ごい）発生，タグ発生など複数の索引情報を付加して記録している。検索処理で必要となるツリー探索を単なるトップダウン処理だけでなく，索引情報を用いてボトムアップ処理や双方向処理を選択的に行うことができ，構造の曖昧（あいまい）表現を含む検索要求を効率的に処理できる。

4　アプリケーション例

われわれのアプローチを適用する分野として，クレーム事例分析，特許分析を考える。

4.1　クレーム事例分析

顧客から得られたクレーム情報のほかに，概念情報を併せ持つ。これらは，すべてXML表現されている。この概念情報を用いた検索が概念検索であり，このアプリケーションは，クレーム情報を対象とした分析環境を提供する。

例えば，「" OS（基本ソフトウェア）"，" 周辺装置 "，などの大きな分類軸で年別傾向を見たい」といった要求がある場合を考える。KFにおいては，過去に実施したクエリを蓄積，再利用することが可能である。そこで " トレンド "などのキーワードを入力して蓄積されている過去のクエリを検索し，自分の要求に近いクエリ（例えば，" 動作 "に関する年別傾向を見たい）を選択する。過去のクエリのパラメータを書き換えることで，自分の観点で再利用する（**図6**）。

更に，「ある年に" 周辺装置 "に関するQ&A件数が変化している」ことに着目したとする。このとき，" 周辺装置 "を展開した" SCSI（Small Computer System Interface）"，" メモリ "，" HDD（ハードディスク装置）"など，一段詳細化したレベルで詳細なトレンドグラフを獲得することができる。

これら一連の分析要求は，「見たい観点で分析を行い」，「見たい部分のデータだけを抽出，統合する」ことの繰返しである。これらの分析要求を表現する言語がKF-QLで，これを処理するのがエンジンである。

4.2　特許分析

特許調査において重要な作業は，関連する特許データを検索して様々な観点から分析し，特許マップを作成することである。従来の特許マップ作成方法は，軸をあらかじめ決定し，それに従い，逐次特許の分析を行うことであり，大きな手間が掛かっていた。しかし，KFを用いることで，オンラインで" 無限の観点 "の特許マップ作成が可能となり，この部分のコストを大幅に減少させることが可能になる。

（注1），（注2），（注3）　Windows，MS-DOS，及びWindowsNTは，米国Microsoft Corporationの米国及びその他の国における登録商標。

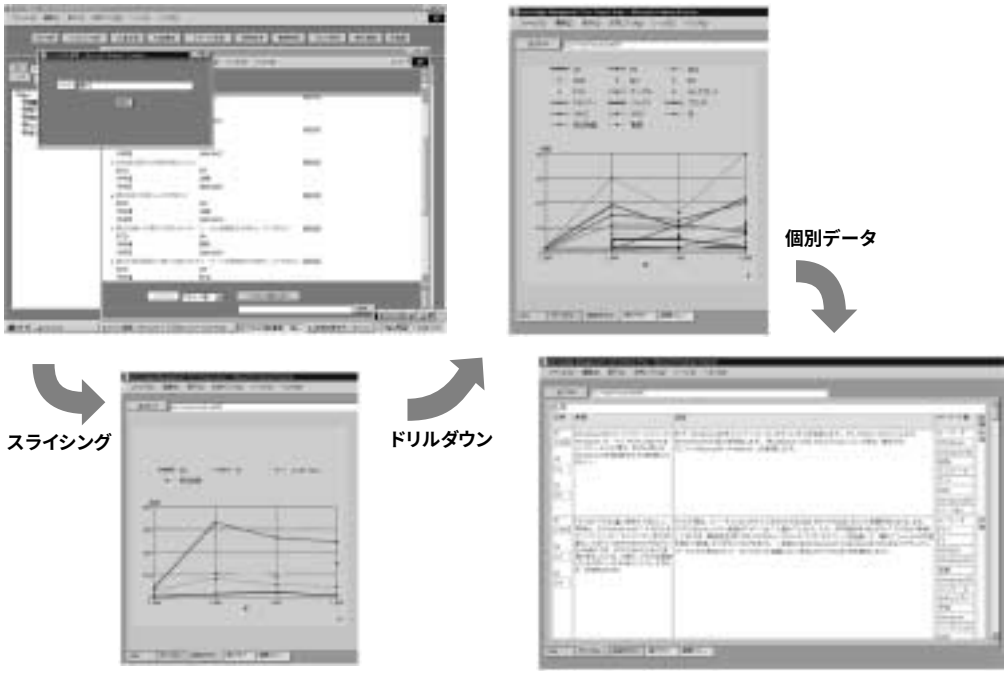

図6．クレーム分析システムの画面例　過去のクエリを再利用して分析を進めていく過程を示している。
Example of user interface for claim analysis application

ユーザーは，分析の軸となる分類名称を入力する。図7では，“機能”×“技術”の軸で特許をマップ化している。

更に，“機能”の一要素である“検索機能”をクリックすると，“検索機能”を展開した特許マップが得られる。

また，各ユーザーが付加した特許に対するランク情報を基に集計を実施したり，各人のコメントリストを一覧表示させたりすることが可能となるなど，情報の共有化を促進できることになり，まさにナレッジマネジメントの目指すべき個人及び組織の創造力の向上につながる。

図7．特許分析でのマップ表示　左上のクエリ再利用ウィンドウにて，“機能”，“技術”というパラメータを入れて実行ボタンを押す。その結果，中央のウィンドウにマップ化された特許群が表示される。
Example of map display for patent analysis application

5 あとがき

XMLデータを構造的に蓄積し，高度に検索・加工できるXML処理エンジンについて述べた。コールセンター事例（p.32参照）の分析支援などの情報分析系のアプリケーションを開発しており，テキスト版OLAP（On-Line Analytical Processing）環境として特徴を出している。

現在，自然言語処理の組込みなどの機能拡張を行うとともに，トランザクション処理を強化している。

また，W3Cにて検索言語の標準化作業が進められているが，策定時に向けて仕様の取込みも並行して進めていく。

文　献

⑴ eXtensible Markup Language（XML）1.0 Specification：http://www．w3.org/TR/REC-xml .
⑵ 服部雅一．“XMLエンジンとナレッジマネージメントシステムについて”．ACM SIGMOD日本支部第15回大会，2000.

服部　雅一　HATTORI Masakazu
研究開発センター 知識メディアラボラトリー研究主務。ナレッジマネジメントシステムの研究・開発に従事。情報処理学会，人工知能学会会員。
Knowledge Media Lab.

末田　直道　SUEDA Naomichi
研究開発センター 知識メディアラボラトリー研究主幹。人工知能，ナレッジマネジメントの研究・開発に従事。情報処理学会，人工知能学会，AAAI会員。
Knowledge Media Lab.