LE-007

複数辞典の鳥瞰が可能な XML 電子辞典システム

XML based Interdictionary Navigation System

小池 勇治† 高野 明彦‡ 絹川 博之†
Yuji Koike Akihiko Takano Hiroshi Kinukawa

1．はじめに

紙ベースの辞典では、用語の意味や概念を調べるのに、次のような問題点がある。

（1）複数の辞典やページの鳥瞰ができない
（2）関連概念を知ることができない
（3）異種文書との類似連想ができない
（4）曖昧語から目的の情報を得るのは難しい

紙ベースの辞典を電子化し、上記問題点を解決することが本研究の目的である。

2．電子辞典システムの具備機能

電子化された辞典システムでは、1.で述べた問題点を解決するために、複数辞典を単に一斉検索する従来方式でなく、辞典項目の類似連想を介して、複数辞典を鳥瞰しつつ、ユーザの興味に応じてナビゲートできることが必要である。

（1）関連概念の検索

関連概念の検索としては、概念体系（シソーラス）を利用する方法と、当該項目の説明文と類似度の高い説明文をもつ項目を動的に連想検索する方法の両方の具備が必要である。また、連想項目や特徴語のユーザ指定に基づき、トピックを絞り込んで検索できる機能の具備が必要である。

（2）辞典と異種文書との高速対応

辞典中の項目の説明文と異種文書との類似連想する機能の具備が必要である。

（3）曖昧用語の特定の支援

利用者が知っている文字列を部分として含む用語一覧を表示することにより、正確な表現の特定を支援することが必要である。

3. 辞典の XML 表現形式

2.を実現するために、複数の辞典を XML 形式で表現し、共通的に扱えるようにする。現在本システムは以下の辞典を記述対象としている。

(1) 岩波情報科学辞典[1]
(a)見出し語とその説明 (b)用語の木

「用語の木」とは、情報科学という分野全体を、一つの木構造で表した概念体系（シソーラス）である。

(2) 岩波ジュニア事典シリーズ（6 種）
(a)見出し語とその説明

(1)、(2)の記述内容を本システムで利用可能な共通形式の辞典 XML データにするためのフォーマットについて述べる。辞書・辞典を XML 電子書籍として表すための形式として DicX[2]というフォーマットが存在する。DicX は項目に関する情報を表現できるが、「用語の木」の表現は対象としていない。そこで DicX に「用語の木」の表現を追加した ex-DicX(extended-DicX)を提案する。

3.1 ex-DicX

ex-DicX は次の 2 つから構成されている。

（1）ex-DicX-term

ex-DicX-term は見出しとその説明の表現用であり DicX に準拠している。その表現構造を図 1 に示す。

（2）ex-DicX-thesaurus

ex-DicX-thesaurus は、「用語の木」の情報のための表現形式であり、項目の子ノードが下位項目となる階層関係として「用語の木」を表現する。（図 2）

3.2 XML 電子辞典の容量

本研究で作成した電子辞典の容量と項目数を表 1 に示す。

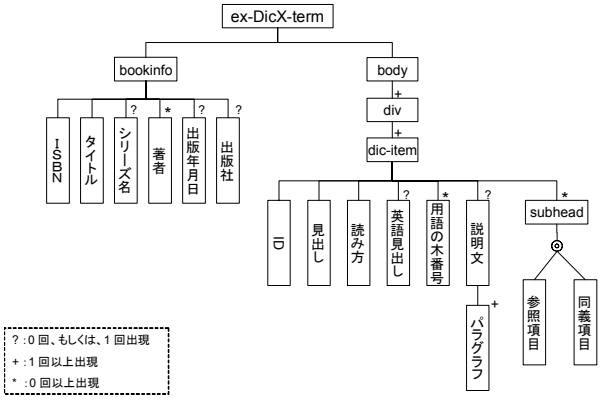

図 1 ex-DicX-term の表現構造

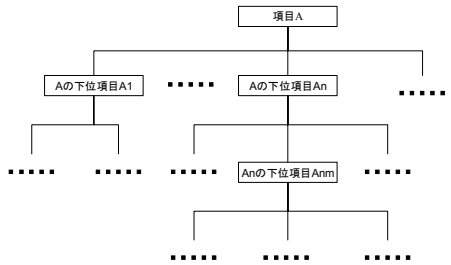

図 2 ex-DicX-thesaurus の表現構造

表 1 電子辞典の容量

辞典名	容量	見出し語数
情報科学辞典	4.93 MB	5468
情報科学辞典(用語の木)	105 KB	―
化学の小事典	451 KB	386
物理の小事典	417 KB	251
地球と宇宙の小事典	455 KB	361
数学の小事典	338 KB	137
化学物質の小事典	353 KB	275
生物の小事典	428 KB	306

† 東京電機大学大学院 工学研究科 情報通信工学専攻
‡ 国立情報学研究所

4. XML 電子辞典システムの機能構成

XML 電子辞典システムは、2.で述べた具備機能のそれぞれ対応した以下の機能を実装しており、そのシステム構成を図 3 に示す。

（1）複数辞典 XML データ制御機能
（2）関連概念検索機能
（3）関連文書検索機能
（4）KWIC 生成機能

4.1 複数辞典 XML データ制御機能

複数の ex-DicX 形式 XML データを一度に扱うことのできる制御機能が必要となる。今回は、表 1 の 7 種の辞典を処理の対象としている。

4.1.1 処理の高速化

XML データはプログラムで扱う際に DOM ツリーという形でメモリに読み込まれる。DOM は一度にすべての文書をメモリに取り込むため、その読み込み時間によって処理速度が遅くなる。そこで要求があるたびにメモリに読み込むのではなく、システム起動時に DOM ツリーをメモリに読み込み常駐させることによって処理の高速化を実現している。

4.1.2 複数 XML データの制御

複数の辞典 XML データの項目情報の部分を取り出しマージして 1 つの DOM ツリーにし、複数の辞典を同時に扱うことを可能にしている。また、マージする際に各項目がどの辞典の情報なのかということを示すために取り出したノードの親ノードにそれぞれの辞典名、ISBN を属性として付加する。

4.2 関連概念検索機能

概念体系検索により、辞典の編著者から提供された静的な体系分類を知ることができ、類似連想検索によりユーザがオンデマンドに必要とする動的な類似関連を得ることができる。

4.2.1 概念体系（シソーラス）を使用する方法

今回概念体系として「岩波情報科学辞典」の「用語の木」を利用した。「岩波情報科学辞典」では、各項目に用語の木の位置を示す木番号が存在する。木番号を基に同じ枝にある項目を関連項目として表示する。

しかしこの方法は、その辞典に対応した概念体系が必要になる。今回利用した「用語の木」は岩波情報科学辞典のみの情報なので、岩波ジュニア辞典シリーズの項目からは利用することができない。

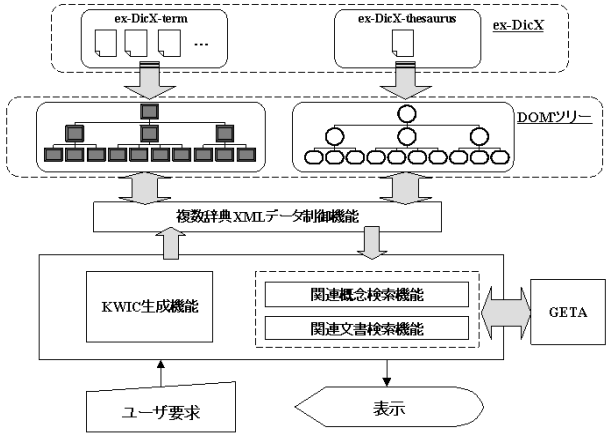

図 3 システム構成

4.2.2 連想計算を利用する方法

指定した項目の見出しと説明文から特徴語を抽出し、その特徴語を基に類似度を計算し、スコアの良い上位 10 件の項目を連想項目とする。この類似度計算の処理に汎用連想計算エンジン(GETA)[3]を利用し、類似度計算に TF,TFIDF,SMART の 3 種類を任意に選ぶことができる。

連想項目の見出しと説明文とから抽出した特徴語をユーザに提示する。上位 10 件の連想項目と特徴語は、トピックをさらに絞り込んだ類似連想検索に使用できる。

4.3 関連文書検索機能

項目と異種文書とを類似度計算し、項目の連想文書を検索する機能である。

4.2.2 と同様に項目の特徴語を抽出する。特徴語と異種文書との類似度を計算し、スコアの良い文書を連想文書とする。ここでも、類似度計算式として TF、TFIDF、SMART 法を指定することができる。

今回は異種文書として毎日新聞 98 年で行い、上位 10 件を連想新聞記事とした。この異種文書の対象を、「新書」「WEB ページ」などに変更することで、異なるジャンルの関連文書を簡易に見つけ出すことが可能である。

4.4 KWIC 生成機能

KWIC とは keyword in context の略で用語の前後の文脈をつけて、用語を中心に配列した形式である。本システムでの KWIC 生成機能では、形態素を単位に前後に文脈を付けることとした。

処理手順は、まず説明文に対して茶筌[4]を用いて形態素解析を行う。用語とマッチした場所から前後に向けて、指定された品詞が連続する限り文字列を取り出す。指定する品詞は任意に変更可能であるが、「名詞、接頭詞、未知語」をデフォルトとしている。文字数で文脈を取り出すよりも、形態素単位で取り出されたほうが利用者は目的の用語を見つけやすくなる。

KWIC 配列を、紙ベースで実現しようとすると膨大なページ数となってしまい、実際には限られた用語に対してしか行うことができない。本システムでは要求に応じてどんな用語に対しても自動的に KWIC 配列を生成する。

5. 利用者インタフェース

利用者インタフェース画面に関し、岩波情報科学辞典の項目「DNA」を例に図 4 に示す。

(1)見出し、英語見出し、出典辞典、読み方、説明文は、ex-DicX-term から取り出した情報である。用語の木番号は 8 桁の英数字であるが「用語の木」と対応させ「情報科学 > 情報基礎 > 生体情報 > 生体の学習の遺伝 > 遺伝と増殖 > 遺伝暗号 > DNA」と表示し、情報科学分野の中での位置付けがわかる。

(2)関連概念検索機能は画面右側に上から「概念体系関連項目」「動的類似連想項目」「連想項目特徴語」が順に表示されている。

「概念体系関連項目」は、本例では「用語の木」における兄弟項目が表示されている。

「動的類似連想項目」は複数辞典の鳥瞰によるものであり、見出し語と出典辞典が表示されている。本例では、「化学の小事典」、「生物の小事典」などの項目にリン

クされた見出し語が表示されている。
「連想項目特徴語」は連想項目の中の特徴語を示している。本例では SMART 法で類似度を計算している。
動的類似連想項目や連想項目特徴語を興味に応じて、選択し、トピック検索を繰り返し実行できる。

(3)画面中央下部の「連想新聞記事」は、関連文書検索機能による。本例では、辞典項目「DNA」に関連の深いヒトゲノムや RNA などの記事が表示されている。

6. おわりに

辞典の紙ベースでの問題点を解決すべく、XML 電子辞典システムを開発した。

(1)複数の辞典を共通的に扱うために XML を採用し、辞典の XML 形式として、DicX を拡張した ex-DicX を新たに提案した。

(2)XML 電子辞典システムには、複数辞典の鳥瞰・ナビゲーションを実装すべく、 (a)関連概念の体系・連想検索 (b)辞典と異種文書との高速対応 (c)曖昧用語の特定の支援、の各機能を具備させた。

(3)岩波情報科学辞典、6 種の岩波ジュニア辞典を実装し、(2)の各具備機能の有効性を確認した。

本システムの機能、使い勝手に関する評価を基にした改良と、より多く辞典の実装が今後の課題である。

謝辞

辞典コンテンツを提供いただいた岩波書店、CD 毎日新聞 98 年版の使用許諾をいただいた毎日新聞社、形態素解析器「茶筌」、汎用連想計算エンジン「GETA」の各開発者の方々、に感謝いたします。

本研究は科学技術振興事業団(JST) 戦略的創造研究推進事業（CREST）プログラムの一環として実施したものです。

参考文献

[1] 長尾真，他: “岩波情報科学辞典.” 岩波書店
[2] イースト株式会社，有限会社ディジタルアシスト: “DicX” http://www.dicx.org/
[3] 高野明彦，他: “汎用連想計算エンジンの開発と大規模文書分析への応用.” http://geta.ex.nii.ac.jp/
[4] 奈良先端科学技術大学院大学自然言語処理学講座: “形態素解析システム茶筌.” http://chasen.aist-nara.ac.jp/

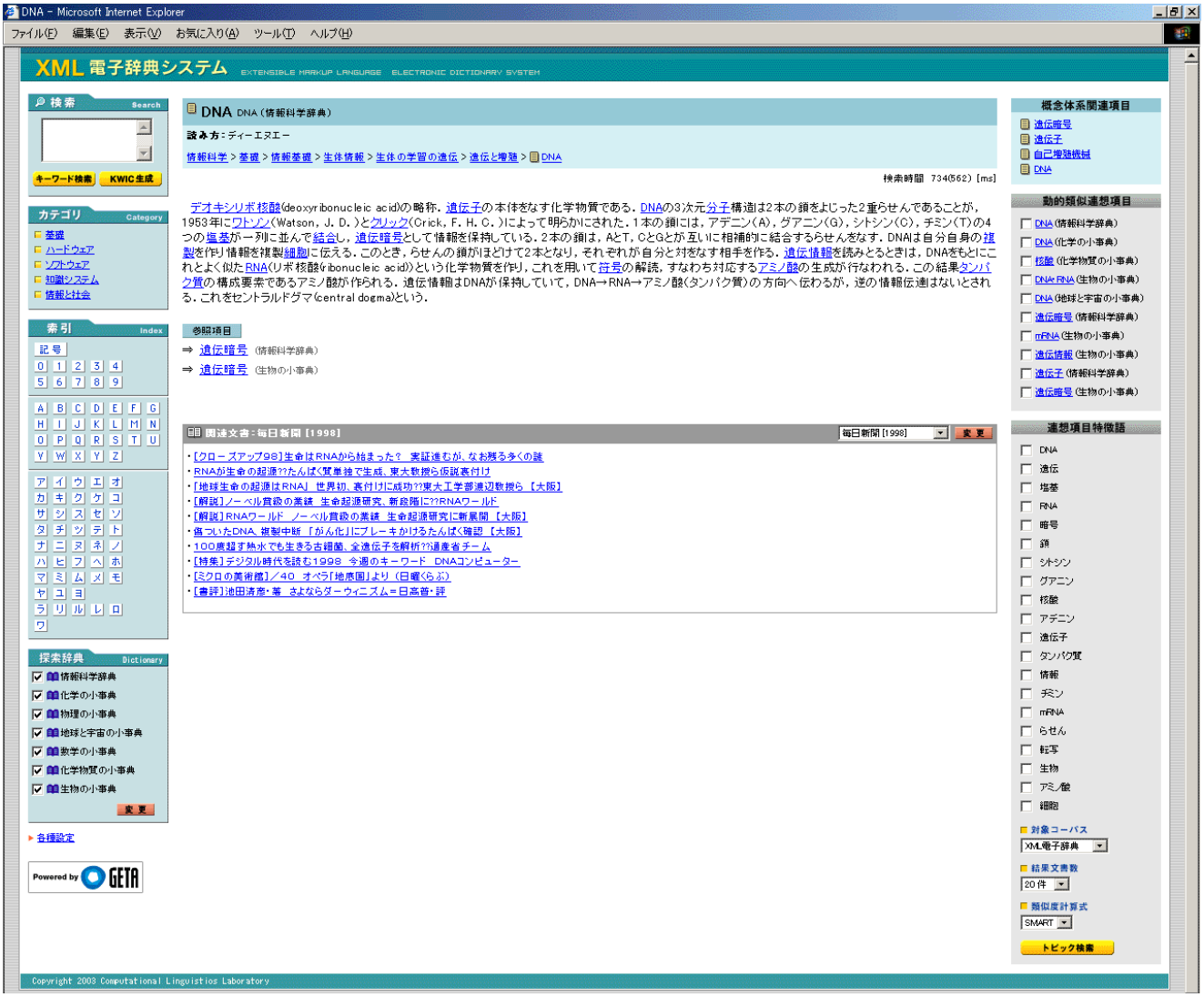

図 4 利用者インタフェース画面：岩波情報科学辞典「DNA」の表示例