関連文書検索システムの開発(2)　－ 構造化文書の処理 －

2U－4

石飛康浩† 永峯猛志† 梅基宏† 倉持勉† 倉橋政之† 増市博†‡ 舘野昌一†
†富士ゼロックス(株) IT 事業開発部 ‡スタンフォード大学 CSLI

1. はじめに

構造化文書を検索対象とする検索システムにおいて、文書中に現われる単語に、その単語が出現する文書構造情報を結合させて１つのトライに登録するインデキシング方法を提案する。

従来のインデキシング方法では、単語が出現する文書構造ごとに複数のテーブルで管理されていた。このような従来のインデックスでは、複数のテーブルに対して単語の検索を行わなければならず、処理アルゴリズムが複雑化してしまうという問題点があった。

筆者らの関連文書検索システムでは、本稿で提案するインデックスを用いることにより、文書構造を利用した検索アルゴリズムを簡易化することができる。

以下、本関連文書検索システムにおける構造化文書の取り扱いについて、構造化文書からの索引語および文書構造情報の抽出方法、インデキシング方法、文書構造を指定した検索方式の３点を中心に報告する。なお、関連文書検索アルゴリズムについては、「関連文書検索システムの開発(4)－関連文書検索－」で報告する。

2. 構造化文書の処理

図１に構造化文書の処理概要を示す。各処理部での処理内容については、文書構造情報の定義、文書構造が付随した索引語の抽出、インデックス生成の３処理に大別して以下に説明する。

なお、本稿で取り扱う構造化文書とは、明示的に文書の構造タグが埋め込まれている文書だけでなく、構造タグと見なすことができるような一定の文字列で表現された暗黙的な構造タグを保持した文書も含める。また、CSV(Comma Separated Value)形式文書についても、各行を検索対象の１単位とし、各カラムが構造を表すものと見なして取り扱う。

2.1 文書構造情報の定義

図１における文書構造定義プロセスでは、図２に示した文書構造定義ファイルを、自動的、あるいは、ユーザを介して生成する。本定義ファイルには、階層構造／構造名、構造コードが記述されている。本ファイルの先頭行には、最上位を示す階層構造「・」および構造名「全体」が定義され、それ以下に実際に存在する階層構造／構造名を記述する。構造コードは、階層構造／構造名をコード化して自動生成される項目である。

本ファイルが自動的に生成可能なものとして、DTD が明確に定義されている SGML 文書群や、CSV 形式の文書などがある。また、ユーザが介在して生成するものとして、定型のプレーンテキストの場合などがある。

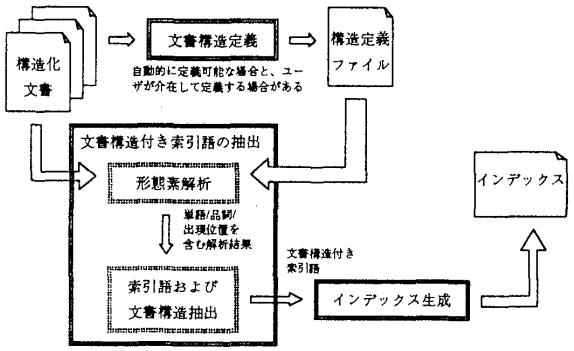
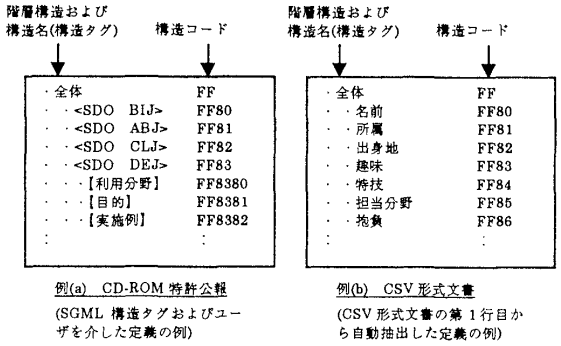


図 1 構造化文書の処理概要



例(a) CD-ROM 特許公報
(SGML 構造タグおよびユーザを介した定義の例)

例(b) CSV 形式文書
(CSV 形式文書の第１行目から自動抽出した定義の例)

図 2 文書構造定義ファイルの例

Relevant Document Retrieval System (2) – Structured Document Processing --
Yasuhiro ISHITOBI†, Takeshi NAGAMINE†, Hiroshi UMEMOTO†, Tsutomu KURAMOCHI†, Masayuki KURAHASHI†, Hiroshi MASUICHI†‡, Masakazu TATENO†
†IT Business Development, Fuji Xerox Co., Ltd. ‡CSLI, Stanford University

2.2 文書構造付き索引語の抽出

前節で定義された文書構造ファイルを用いて、検索対象文書(群)を形態素解析[1]する。本システムが採用している形態素解析ツールでは、結果として単語/品詞/構造名が得られる。この結果から助詞や助動詞を除く自立語を抽出し、その構造名を構造コードに変換して、単語の後方に結合させたものを索引語とする。構造コードを含む索引語は、検索単位の識別番号(以下、UID)と共にインデックス生成用のデータとして一時保持される。なお、現在、本システムでは、文字コード EUC をベースとした処理を行っており、図 2 に例示した各構造コードは 16 進コードで索引語に付加されている。

2.3 インデックスの生成

検索対象となる全ての構造化文書について、構造コード付きの索引語および UID の対を抽出した後、トライ構造形式のインデックスを生成する。本インデックスは、索引語の先頭文字列から構造コードまでが順にインデックスに登録され、終端に UID テーブルへのポインタが記述されているので、文書構造を指定した単語の完全一致および前方一致は、容易に結果が得られる仕組みになっている。

3. 構造化文書の検索方法

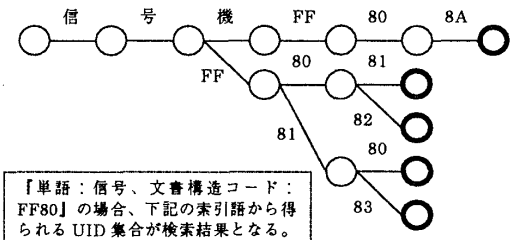
本システムでは、構造コード付きの索引語が登録されたインデックスを用いることにより、文書構造を指定した単語の検索が容易に行える。本インデックスを用いた単語の完全一致検索、前方一致検索の方法を説明する。図 3 に各検索事例を示す。

3.1 完全一致検索

ユーザが文書構造を指定し、さらに、単語の完全一致検索を指定した場合、単語の後方に構造コードを付加した文字列を用いて、本インデックスに対して前方一致検索を行う。その結果得られた索引語リストから検索結果となる UID リストを求める。本インデックスに対して、前方一致検索を行うことにより、ユーザが指定した文書構造の下位構造も含めた検索が可能になる。

3.2 前方一致検索

ユーザが文書構造を指定し、さらに、単語の前方一致検索を指定した場合、最初に、単語のみを用いて、本インデックスに対して前方一致検索を行う。その結果得られた索引語リストの中から、ユーザが指定した構造コードを含む索引語のみを抽出し、その索引語から検索結果となる UID リストを求める。単語の前方一致検索をした場合でも、ユーザが指定した文書構造の下位構造を含めた検索が可能になる。



「単語：信号、文書構造コード：FF80」の場合、下記の索引語から得られる UID 集合が検索結果となる。

完全一致検索
信号 FF8081, 信号 FF8082
前方一致検索
信号 FF8081, 信号 FF8082,
信号機 FF808A

◎は終端ノードをあらわす。

図 3 インデックスへのアクセス例

4. おわりに

構造化文書のように階層的な文書構造をもった検索対象文書に対して、単語の後方に文書構造を表現した構造コードを結合した索引語をトライに登録したインデックスの生成およびアクセス手法を提案した。本手法によるインデックスを用いることにより、簡易なアルゴリズムで文書構造を指定した単語の検索が可能になった。

参考文献

[1] 増市,山浦,小山,舘野,「形態素解析を用いた全文検索システムとその応用」, 情報処理学会自然言語処理, 102-3, pp.17-24(1994.7)