1S－5　製品情報広域検索システムにおけるデータベース自動構築方式

森口 修　今村 誠　鈴木 克志
三菱電機株式会社　情報技術総合研究所　音声・言語インタフェース技術部

１．はじめに

CALS(Commerce At Light Speed)の普及に伴い、広域ネットワーク上のサーバにおいて、企業の製品広告などの情報公開が盛んになってきた。膨大な量の製品カタログがネットワーク上に散在する状況では、製品情報の検索は非常に重要な機能となる。

本稿では、まず製品情報の検索に関する要求機能を分析し、製品カタログのSGML(Standard Generalized Markup Language)文書形式を設計する。次にパラメタ検索、分類検索、キーワード検索という３種類の検索方式で必要となる索引の自動作成が可能な製品情報広域検索システムについて述べる。

２．製品検索機能の要求分析

製品の検索機能に関する要求は、検索方式と索引自動作成の２つがある。製品の検索方式は、製品カタログに記述される情報の内、何を検索キーとするかによって表１に示すような３種類に分ける。

表１．製品の検索方式と検索キー

検索方式	検索キー
パラメタ検索	製品が有する特性からなる製品パラメタ
分類検索	製品がどの分野に属するかという分類情報
キーワード検索	製品を代表するキーワード

索引は、検索キーと製品カタログとを関連付けたデータであり、検索キーから製品を高速に検索することを目的として、あらかじめ作成しておく必要がある。ただし、製品カタログが大量かつ増加し続けるという状況では、索引を自動的に作成することが要求される。索引を自動作成するには製品カタログから検索キーを自動抽出する機能が必要である。検索キーの内、製品の分類情報は、利用者毎に使い易い分類体系が異なる点や、将来現れる新たな製品分類をあらかじめ決めることができないため、製品カタログの内容から自動判別する必要がある。

A Method of Automatic Database Creation in the Product Information Retrieval System on Wide-area Network
Osamu Moriguchi, Makoto Imamura, Katsushi Suzuki
Human Media Technology Dept. Information Technology R&D Center
Mitsubishi Electric Corporation.
5-1-1 Ofuna, Kamakura, Kanagawa 247, Japan

３．製品カタログＳＧＭＬ文書形式の設計

前章で述べた各検索方式の検索キーを製品カタログから自動抽出することを目的とした DTD (Document Type Definition)を定義することでSGML文書形式を設計する。

パラメタ検索の検索キーとなる製品パラメタは、機械的に正確に抽出することが必要であるため、機械可読なタグを付けた形式の論理構造とする。例えば<特性>タグの繰り返しによるリスト構造、及び<特性名>、<特性値>の対構造により、複数の製品パラメタの名称および値を記述する。

分類検索の検索キーとなる製品の分類情報は、製品カタログの内容から自動判別することにより、製品カタログに直接記述しなくても良いようにする。

キーワード検索の検索キーとなるキーワードは、概要や特徴などの製品を説明するテキスト部分から抽出する。例えば、<製品説明>というタグによって、抽出対象のカタログの説明テキストの範囲を機械的に判別可能とする。

４．製品検索のための索引自動作成

今回試作した製品情報広域検索システムにおいて、パラメタ検索、分類検索、キーワード検索に必要とされる索引を自動作成するため、各検索方式の検索キーを製品カタログから自動抽出する方法について述べる。図１にシステム構成を示す。

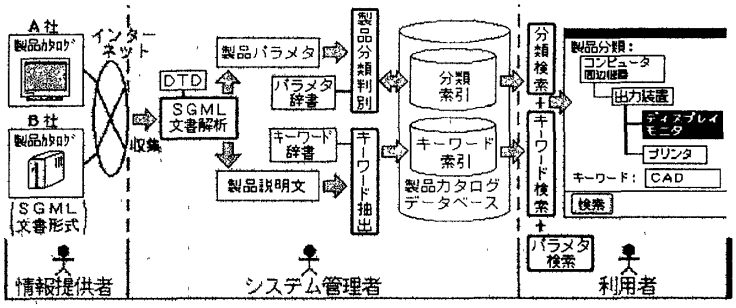

図１．製品情報広域検索システムの構成

４．１　製品パラメタ情報の抽出

製品パラメタは、SGML形式で記述された製品カタログの論理構造を解析したパーザの出力結果から抽出する。ただし、製品パラメタの抽出処理はDTD毎に記述する必要がある。そこで、図２に示すような変数を含むパター

ン記述による製品パラメタ抽出ルールを導入する。パラメタ抽出ルールで表わされる文書構造パターンと製品カタログの任意の部分とのパターンマッチを実行し、変数とマッチした箇所を製品パラメタの名称および値の対として抽出する。

```
パラメタ抽出ルール:(特性 (特性名 $PARAM_NAME)
                        (特性値 $PARAM_VAL  ))
製品カタログの一部:  <特性><特性名>画面サイズ
                        <特性値>17インチ
抽出される製品パラメタ:名称= 画面サイズ
                        値 = 17インチ
```

図2．製品パラメタ抽出ルールと抽出例

また、同一のパラメタであってもメーカ毎に製品パラメタの名称が異なることがある。そこで、図3に示すようなパラメタ辞書を用いてＩＤが同一となる同義語や異表記語を同一のパラメタとみなして統一して扱うことにより、製品パラメタの抽出精度を向上させる。

ＩＤ	パラメタ名称		
MONSIZ	画面サイズ	モニタサイズ	
CASHMEM	キャッシュ	キャッシュメモリ	キャッシュ容量
CONNECT	コネクタ	コネクタ形状	
PRNMET	プリント方式	印刷方式	印字方式
SUPOW	電源	電源入力	入力電源
：	：	：	：

図3．パラメタ辞書

4．2 製品分類情報の抽出

従来から、文書の特徴量を表わすベクトルを文書中の単語の出現頻度を元に統計的に作成し、ベクトル間の距離を文書間の類似度とみなすという方法が用いられている[1]。本システムでは、単語ではなく4．1で抽出した製品パラメタによって特徴ベクトルを作成する。この理由は、製品パラメタは製品カタログの必須項目であるだけでなく、分類が異なれば使用される製品パラメタが異なるからである。

製品パラメタによる特徴ベクトルの例を図4に示す。横軸は製品パラメタ、縦軸は新着の製品カタログおよび既存の製品分類の特徴ベクトルである。新着の製品カタログは「画面サイズ」や「ドットピッチ」などのディスプレイモニタ特有の製品パラメタを多く有するため、特徴ベクトルを比較した結果、分類先は「ディスプレイモニタ」と判別される。また、特徴ベクトルの学習データに「光磁気ディスク装置」の製品カタログが含まれていない場合でも、「容量」や「平均シーク時間」などのディスク装置特有の製品パラメタによって、分類先を「ディスク装置」と判別することができる。

本機能によれば、利用者に応じた製品分類毎に製品カタログのサンプルをあらかじめ用意しておけば、新着の製品カタログを自動分類できるので、利用者毎の分類検索が可能になる。

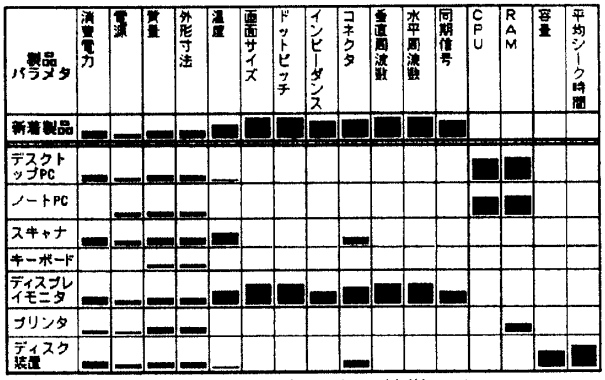

図4．製品パラメタによる特徴ベクトル

4．3 製品キーワード情報の抽出

キーワード付けの作業コスト、付与するキーワードのばらつきが問題とされるため、従来から語の表現形式や構造を利用する解析的手法や語の出現頻度を利用する統計的手法によってテキストからキーワードを自動抽出する研究がなされている[2]。

本システムでは統制キーワード辞書を用いて製品カタログからキーワードを抽出する方式に加えて、語の出現パターンからキーワードを判別する方法を導入する。例えば、「○○用」、「○○に最適」といった製品カタログに多く見られる語の出現パターンをとらえることにより、「用途」というカテゴリのキーワードを抽出する。

5．まとめ

広域ネットワーク上の製品カタログを検索するシステムを試作した。本システムの特長は、収集した製品カタログをパラメタ検索、分類検索、キーワード検索するための索引を自動作成することである。特に、製品パラメタの抽出をSGML文書の構造解析によって行なったこと及び製品分類を製品パラメタに基づいて判別したことにより、抽出および分類の精度を高めることができた。

今後の課題は、キーワード抽出精度の向上のための語の出現パターンの記述能力の向上と、キーワード抽出用辞書の作成を支援するために単語や語の出現パターンを文書から自動抽出することである。

参考文献

[1] 河合：意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，Vol.33, No.9, pp.1112-1122(1992)

[2] 諸橋：自動索引付け研究の動向，情報処理，Vol.25, No.9, pp.918-925(1984)