デジタル・ドキュメント　5－4
（1997．1．31）

製品情報広域検索システムにおける検索方式

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

〒247 神奈川県鎌倉市大船 5-1-1

ＣＡＬＳ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文書型定義を設計する。次に、分類検索、パラメタ検索、キーワード検索の３種類の検索に必要な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。このシステムは、ＳＧＭＬ文書を構造解析し、製品の特性を製品パラメタとして抽出する。製品パラメタで製品情報を特徴付けることにより、分類検索における製品分類の判別精度を高めることができた。

A Product Information Retrieval System on the Wide-area Network

Osamu Moriguchi　　Makoto Imamura　　Katsushi Suzuki

Information Technology R&D Center
Mitsubishi Electric Corporation

5-1-1 Ofuna, Kamakura, Kanagawa 247, JAPAN

As popularization of CALS, advertisements of product information are published on the enterprise's servers through the wide-area network. Under the circumstance that huge amount of the product catalogs are distributed on the network, the product information retrieval is one of the most important tools. Firstly, we investigate the requirements of product information retrieval and design an DTD of SGML for the product catalogs. Next, we describe the wide-area information retrieval system for the product catalogs. This system automatically generates indexes for category-based retrieval, parametric retrieval and controlled keyword retrieval. High precision of the parameter extraction and the product discrimination results from the product parameters extracted from SGML documents and product categorization based on the product parameters.

1．はじめに

ＣＡＬＳ[1]の普及に伴い、広域ネットワーク上の企業のサーバにおいて、企業広告や製品宣伝などの情報が公開されるようになる。このような膨大な量の製品カタログがネットワーク上に散在するという状況においては、高精度な製品情報の検索は非常に重要な機能となる。製品検索機能は、製品の利用者だけでなく、製品の提供者の要求でもある。すなわち、利用者にとっては必要な製品だけが漏れなく検索できることが求められ、提供者にとってはその製品を必要としている利用者にだけ確実に届くことが求められる。このように、製品の提供者および利用者の双方において製品カタログの利用目的が一致するにもかかわらず、文書内容・形式の多様性、文書所在の分散性、文書量の大規模性という問題が、製品検索機能の向上を阻む要因となっている。

既にインターネットを介してWWW[2]により参照可能なＨＴＭＬ[3]形式の文書は膨大な量に達し、それらを検索するサイトも多数出現している。しかし、ＨＴＭＬは表示機能およびハイパーリンク機能を主な目的として設計されたＳＧＭＬ[4]文書形式の一例である。ＳＧＭＬの本来の仕様は、文書の表示やハイパーリンクだけでなく、検索や交換などの文書の様々な利用目的を考慮した文書の論理構造を定義することが可能なメタ言語である。言い換えると、文書の利用目的を考慮して設計したＳＧＭＬ文書は、文書の利用のために必要な論理構造が機械可読となるということである。

本稿では、まず製品検索の要求機能を分析し、製品カタログのＳＧＭＬ文書形式を設計し、広域ネットワーク上に検索を目的として公開される製品カタログの形式を提案する。次に、製品固有の特性からなる製品パラメタが製品の自動分類に有効であること、製品カタログの形式を構造化することにより製品パラメタの抽出精度を高めることができることを示す。

2．製品検索機能の要求分析

製品検索機能に対する要求は、検索方式と索引自動作成の２つがある。

製品検索方式は、製品カタログが含むかまたは付与する情報の内、どの情報を検索キーとするかである。ここでは、分類検索、パラメタ検索、キーワード検索の３つの検索方式について検討する。これらの３つの検索方式はそれぞれ、製品がどの分野に属するかという分類情報、製品固有の特性である製品パラメタ情報、製品を代表するキーワード情報を検索キーとする。

索引は、検索キーと製品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、広域検索システムにおいては、製品が大量かつ増加し続けるという状況であるため、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となる。

まず、３つの検索方式の説明と、それぞれの検索方式における検索キーを製品カタログから自動抽出する方針について述べる。

---

[1] CALS:Commerce At Light Speed

[2] WWW:World Wide Web

[3] HTML:HyperText Markup Language

[4] SGML:Standard Generalized Markup Language

2．1 分類検索

分類検索は、製品がどの分野に属するかという分類情報を検索キーとし、検索時には製品の分類を対話的に指定することにより製品カタログを絞りこむ検索方式である。製品の分類を階層関係からなる体系としておけば、分類検索のインタフェースは階層木のブラウジングとすることができる。

例えば、図1に示すように「プリンタ」という分類を検索条件として指定すると、「プリンタ」に分類された製品カタログだけが検索される。

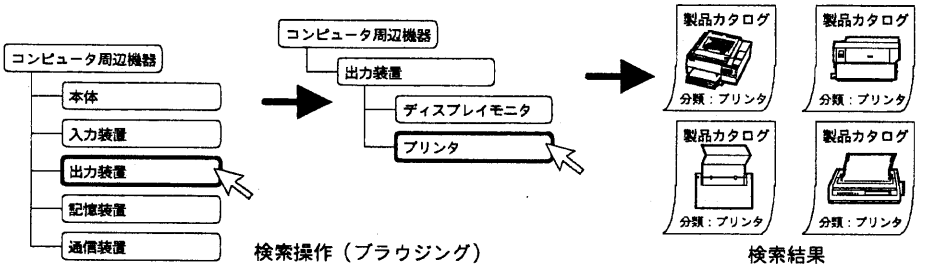


図1．分類検索の例

製品の分類は、提供者が指定する場合と利用者が指定する場合がある。

提供者が指定する分類は標準化されたものであり、分類の名称または識別子が製品カタログに記述されるはずであるから、分類体系および分類の記述形式の標準化という問題に帰着するため、ここでの議論の対象とはしない。

利用者が指定する分類は、利用者が検索し易いように定義するものであるから、利用者毎に分類体系が異なる。したがって、製品の分類情報を製品カタログから自動抽出するためには、利用者が定義した分類を学習する機能と、新規に収集した製品の分類を自動判別する機能が必要である。分類の学習は、分類済みの製品の特徴を記憶する機能であり、分類の自動判別は未分類の製品の特徴を分類済みの製品の特徴と比較する機能である。製品特徴の算出には、次節に述べる製品パラメタを用いる。

2．2 パラメタ検索

パラメタ検索は、製品が有する特性を検索キーとし、検索時には特性の値を指定することにより製品カタログを絞りこむ検索方式である。製品が有する特性をここでは製品パラメタと呼ぶ。

例えば、図2に示すようにプリンタの製品カタログであれば、製品パラメタは印刷速度や解像度などである。製品パラメタは製品の分類によって偏るため、前述の分類検索によって製品を絞り込んだ後にパラメタ検索を実行するという手順が一般的である。これを逆に考えると、同一の分類に属する製品は共通の製品パラメタを多く有するということであり、製品パラメタは製品の特徴を表わすのに有効であるといえる。

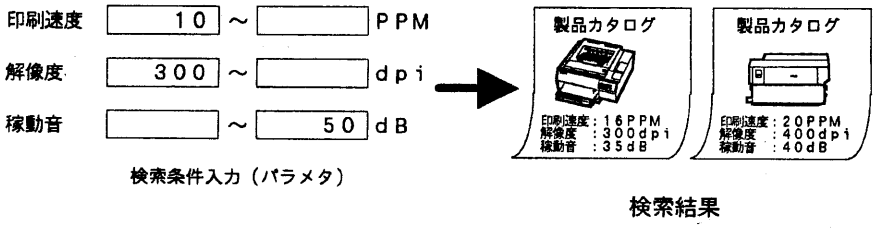


図2．パラメタ検索の例

製品パラメタの自動抽出は正確さが要求されるため、ＳＧＭＬ形式で記述された製品カタログの文書構造を解析することにより行なう。このため、製品カタログ中に記述される製品パラメタは、製品パラメタの名称および値をＳＧＭＬの機械可読なタグ[5]で明確に区切る。

２．３ キーワード検索

キーワード検索は、製品カタログ中に付与されたキーワードを検索キーとし、検索時にはキーワードを指定することにより製品を絞りこむ検索方式である。

例えば、図３に示すように「アウトラインフォント」というキーワードを検索条件として指定すると、「アウトラインフォント」というキーワードが付与された製品カタログだけが検索される。

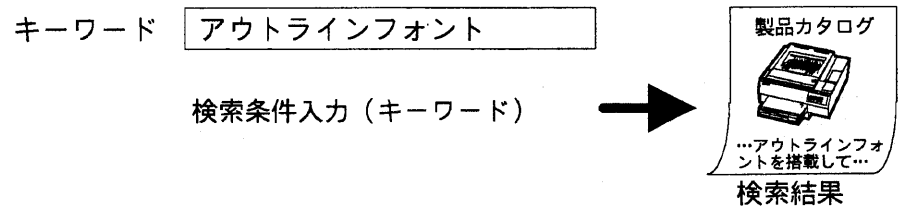


図３．キーワード検索の例

キーワードの自動抽出は、概要や特徴などの製品を説明するテキスト部分を対象として行なう。このため、製品カタログ内の抽出対象のテキストはＳＧＭＬの機械可読なタグによって明確に指定する。

３．製品カタログのＳＧＭＬ文書形式の設計

２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方針に沿って、ＤＴＤ[6]を定義することにより製品カタログのＳＧＭＬ文書形式を設計する。論理構造および文書サンプルは図４（ａ），（ｂ）のようになる。

分類検索の検索キーとなる製品の分類情報は、製品パラメタを基に自動判別するため、製品カタログ中に分類情報を記述する要素は特に設けない。

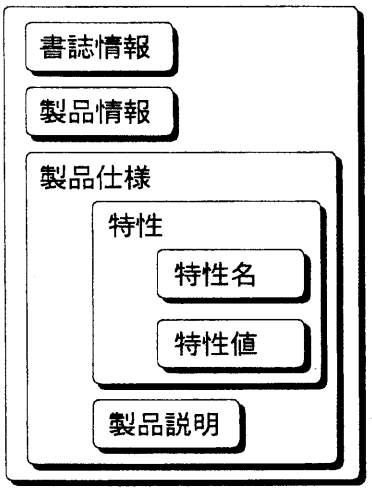
パラメタ検索の検索キーとなる製品パラメタは、機械的に正確に抽出することが必要であるため、機械可読なタグによって詳細な形式の論理構造とする。図４の例では、１つの＜特性＞タグで表わされる要素によって１つの製品パラメタを記述し、＜製品仕様＞という要素の下位に＜特性＞要素を繰り返すリスト構造によって複数の製品パラメタを記述している。＜特性＞は＜特性名＞と＜特性値＞という２つの下位要素から成り、＜特性名＞には製品パラメタの名称を、＜特性値＞には製品パラメタの値を記述している。したがって、製品パラメタの名称および値は、これらのタグで表わされる要素の論理構造を解析することによって正確に自動抽出できる。

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。図４の例では、＜製品説明＞というタグによって、抽出対象の説明テキストの範囲が機械的に判別可能である。ただし、説明テキストの下位の論理構造に対しては特に形式的な制約を課さない。

---

[5] タグ:SGML における、論理構造要素(ELEMENT)を指定する印。名前と属性が付与される。

[6] DTD:Document Type Definition（文書型定義）




```
<!DOCTYPE 製品カタログ SYSTEM "product.dtd">
<製品カタログ>
<文書情報>
<文書名>ＲＤシリーズ製品カタログ（ＲＤ１７Ｇ２）
<著者>○○株式会社
<発行日><年>95<月>12
</文書情報>
<製品情報>
<製品分類>コンピュータ周辺機器
<製品名>カラーディスプレイモニタ
<型番>ＲＤ１７Ｇ２
<メーカー>○○株式会社
<販売元>○○映像情報デバイス事業部
<価格>１９０，０００円
</製品情報>
<製品仕様>
<特性><特性名>画面サイズ<特性値>１７インチ</特性>
<特性><特性名>ＡＧピッチ<特性値>０．２５ｍｍ</特性>
<特性><特性名>水平走査周波数<特性値>２４ｋＨｚ～８６ｋＨｚ</特性>
                :
<特性><特性名>質量<特性値>２１．５ｋｇ</特性>
<製品説明>
<概要>
<段落>きわだつ表現性能。明るさとコントラストが違います。</段落>
<段落>シャドウマスクに比べ、蛍光面の開口部が大幅にアップした、
アパーチャーグリルを採用しました。だから明るさと白さが違います。
３次元グラフィックスやＣＧ／ＣＡＤに最適です。</段落>
<箇条書き>
<項目>プラグ＆プレイ機能対応
<項目>２４～８４ｋＨｚのワイドレンジマルチスキャン
<項目>ＶＣＣＩ－２種、ＭＰＲ２、国際エネルギースタープログラム準拠
</箇条書き>
</概要>
                :
</製品説明>
</製品カタログ>
```

（ａ）論理構造定義　　（ｂ）文書サンプル

図４．製品カタログの論理構造定義と文書サンプル

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、分類検索、パラメタ検索、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に、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５にシステム全体の構成を示す。

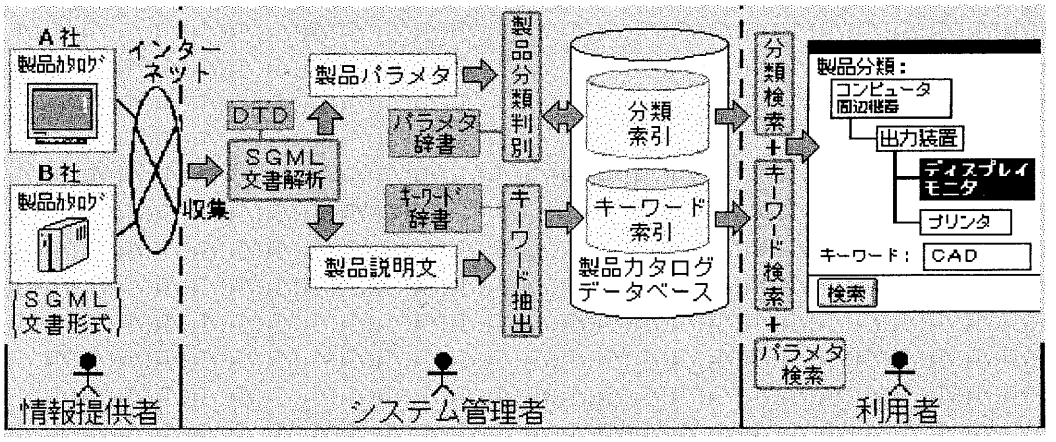


図５．製品情報広域検索システムの構成

４．１ 製品パラメタ情報の抽出

製品分類は製品パラメタを基に判別するため、先に製品パラメタの抽出について述べる。

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することにより抽出する。製品パラメタを記述する部分にどのようなタグを挿入するのかは製品カタログの論理構造定義によって異なるため、ＤＴＤ毎にパラメタを抽出する処理を記述する必要がある。

そこで、図６に示すような“（”で始まる開始タグ、“）”で始まる終了タグ、“$”で始まる変数からなるパターン記述によるパラメタ抽出ルールを導入する。パラメタ抽出ルールで表わされる文書構造パターンと製品カタログの任意の部分とのパターンマッチを実行し、変数“$PARAM_NAME”および“$PARAM_VAL”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　）特性名
                        （特性値　$PARAM_VAL　　）特性値
                    ）特性
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
                    ＜／特性＞
抽出される製品パラメタ：名称　　＝画面サイズ
                        値　　　＝１７インチ
```

図６．パラメタ抽出ルールとパラメタ抽出例

また、同一の製品パラメタであっても製品パラメタの名称が異なる可能性がある。そこで、図７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製品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			
TEMP	温度	温度条件	環境温度	使用周囲温度	動作温度
AMOS	湿度	湿度条件	環境湿度	使用周囲湿度	動作湿度
SUPOW	電源	電源入力	入力電源		
REDDEN	読み取り解像度	読取解像度	読み取り密度	読取密度	
：	：	：	：	：	：

図７．パラメタ辞書

４．２ 製品分類情報の抽出

製品の分類は、製品パラメタを基に自動判別する。従来から、類似する文書は同一の単語もしくは文字を同様の頻度で含んでいるという前提により、文書の特徴量を表わすベクトルを文書中の単語や文字の出現頻度を元に統計的手法により作成し、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]。この手法はテキスト情報のみからなる新聞記事などの文書の分類に用いられるが、製品パラメタが抽出可能な製品カタログの場合、単語の代りに製品パラメタによって特徴ベクトルを作成する方法が有効である。なぜなら、製品パラメタは形式の解析により正確に抽出することが可能であり、さらに出現頻度は分類に依存して１または０であり曖昧性がない。

製品パラメタによって算出した特徴ベクトルを図８に示す。図８において、横軸は製品パラメタ、縦軸は新着製品および製品分類、内部の棒グラフの高さは特徴ベクトルの各成分の値である。特徴ベクトルの各成分の値は、製品パラメタの頻度からχ二乗値[2]により重み付けし、長さを１に正規化することにより算出した。新着の製品カタログは「画面サイズ」、「ドットピッチ」などのディスプレイモニタにのみ偏って出現する製品パラメタを多く有し、図９に示すようにディスプレイモニタとの類似性が最も高くなり、分類先をディスプレイモニタと判別できる。

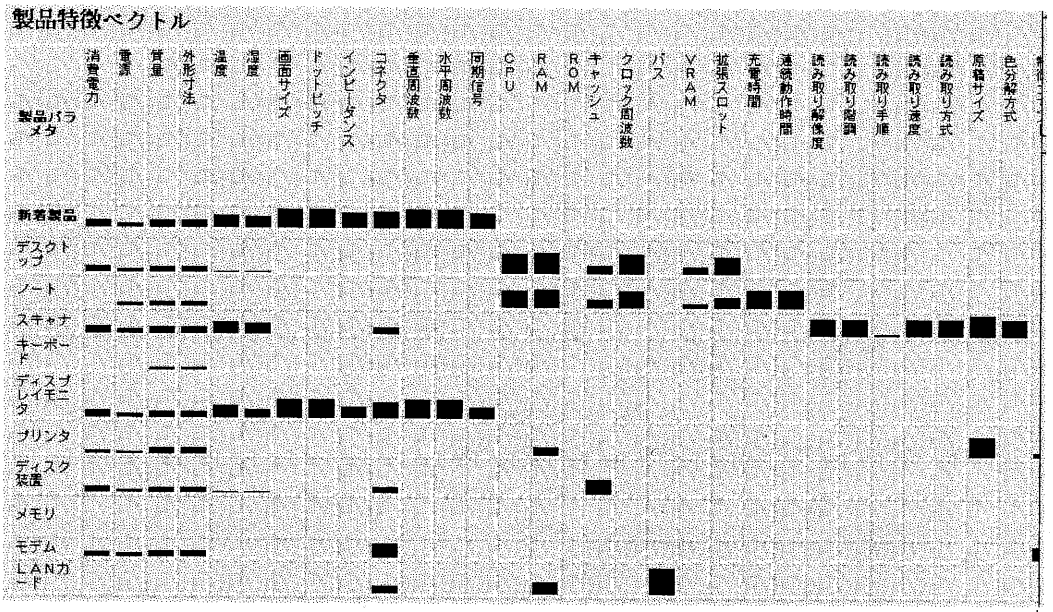


図８．製品パラメタによる特徴ベクトル

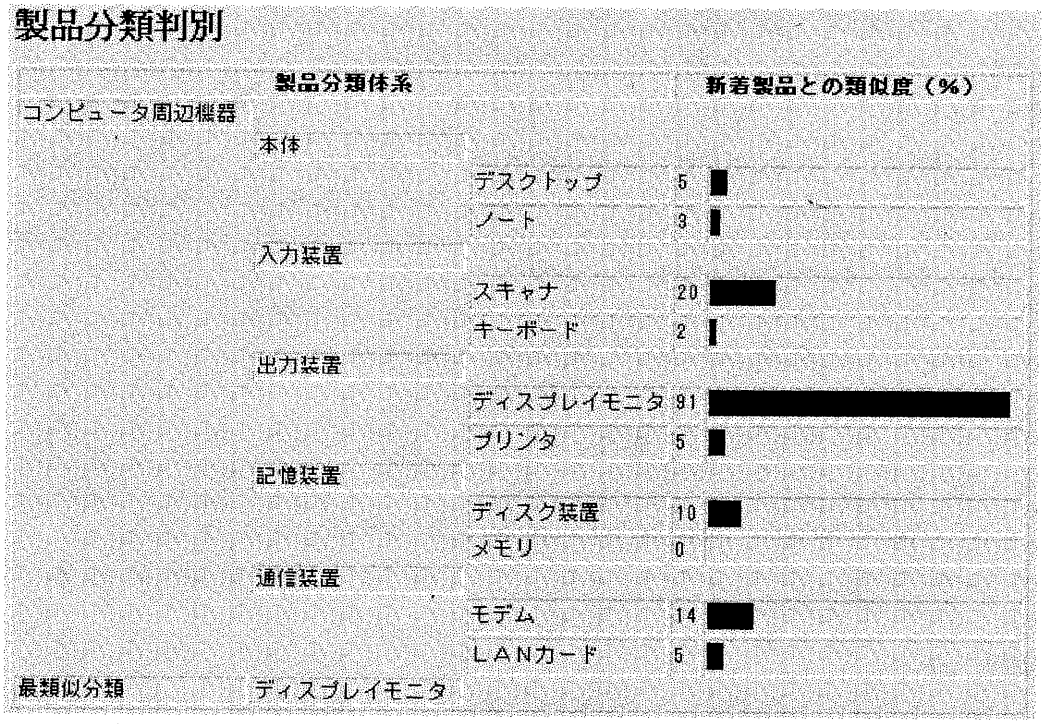


図９．特徴ベクトルの比較

４．３ 製品キーワード情報の抽出

ＳＧＭＬ文書においては、論理構造要素としてキーワードを文書中に作成者が記述することが可能であるが、人手によるキーワード付けの作業コスト、付与するキーワードの作業者によるばらつ

きが問題とされるため、語の表現形式や構造を利用する解析的手法[3][4]や語の出現頻度を利用した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[5]。

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統制キーワード辞書に登録されている語をキーワードとして抽出する方式に加え、語の表現形式を利用した解析的手法として形態素列中の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、「用途」というカテゴリを持つキーワードは「○○用」、「○○に適した」、「○○に最適」の○○のように接尾語「用」、動詞「適する」、形容詞「最適」の前に出現する。このような語の出現パターンをとらえることにより、キーワードを抽出すると同時にキーワードにカテゴリを付与する。図１０は語の出現パターンをとらえるテキストパターン部とキーワードにカテゴリを付与するカテゴリ抽出部からなるキーワード抽出ルールの例である。

キーワードをカテゴリによって意味分類することで、キーワード検索時にキーワード候補をカテゴリ毎に分類して検索者に提示することが可能となる。

テキストパターン部	カテゴリ抽出部
[$1 名詞][**用** 接尾語]	用途：$1
[$1 名詞][に 格助詞][**最適** 形容詞]	用途：$1
[$1 名詞][が 格助詞][**可能** 名詞]	機能：$1
[$1 名詞][を 格助詞][**内蔵** サ変]	付加機能：$1
[$1 名詞][に 格助詞][**準拠** サ変]	規格：$1
[$1 名詞][を 格助詞][**採用** サ変]	方式：$1

図１０．キーワード抽出ルール

５．まとめ

広域ネットワーク上のＳＧＭＬ化された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、広域ネットワークから収集した製品カタログを分類検索、パラメタ検索、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、以下の２点により製品パラメタの抽出精度および製品分類の判別精度を高めることができた。

(1) ＳＧＭＬ文書を構造解析し、製品の特性を製品パラメタとして抽出する。

(2) 製品パラメタで製品情報を特徴付け、製品分類を判別する。

キーワード抽出に関する以下の２点は今後の課題である。

(1) キーワード抽出対象のテキスト範囲はＳＧＭＬにより明確化できるが、テキストの形式に制約を課すことはできないため、辞書にないキーワードや抽出ルールにマッチしない表現のキーワードが抽出できないという問題（辞書の網羅性）があった。これには辞書や抽出パターンを単語や単語の共起頻度などにより自動的に学習して拡充するなどの仕組みが今後必要となる。

(2) 辞書の検索およびパターンマッチによるキーワード抽出では、同義語、複合語の範囲などの同定（対象の曖昧性の解消）が課題である。

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] 田村、渡辺、原、笠原：統計的手法による文書自動分類，第３６回情報処理学会全国大会論文集，pp.1305-1308(1988)
[3] 松尾、木本：抽出パターンの階層的照合に基づく日本語テキストからの内容抽出，情報処理学会論文誌，Vol.36, No.8, pp.1838-1844(1995)
[4] 小川、望主、別所：複合語キーワードの自動抽出法，情報処理学会自然言語処理研究会，97-15(1993)
[5] 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)