F-010

WWWを利用した企業情報比較支援システムの開発
Development of Company Information Analysis System Using WWW Contents

大沼 宏行† 松平 正樹† 上田 俊夫† 渕上 正睦† 森田 幸伯†
Hiroyuki OHNUMA Masaki MATSUDAIRA Toshio UEDA Masachika FUCHIGAMI Yukihiko MORITA

1．はじめに

インターネット上では，企業情報や製品情報を集めた様々なサイトが存在している．例えば，製品情報に関するサービスとして，製品の価格情報等を管理し，ユーザが興味のある製品を価格等で比較することができるサイトがある．また，製品情報の抽出に関する研究として[1][2]があり，嶋田ら[1]は，製品情報を記載したWebページからスペック情報を抽出し，製品間のスペック情報の比較によって最上位機種の推定などを可能にしている．一方，徳永ら[2]は，新製品情報を記載したニュース記事から製品の特徴情報を抽出している．

これらの研究は，製品の購買者が，特定の製品カテゴリ内で製品間の違いを比較検討するのに特に有効である．しかし，新商品の企画全般を支援する目的には十分ではない．企画全般を支援するためには，企業ごとの製品ラインナップや製品の技術的な特徴等を網羅して，ユーザに提示することが必要である．

我々は，新商品の企画全般を支援するシステムを開発することを目標としている．すなわち，次の機能をもつシステムである．

- 自社や各企業がどのようなカテゴリの製品を開発しているのかを網羅的に俯瞰できる．
- 製品情報とプレスリリース情報，企業情報など各種情報とを関連づける．
- プレスリリースから製品の特徴を示す単語を抽出する．

このうち，各種情報との関連づけについては，新商品企画支援という目的に限定していないが，松平ら[3]によって研究されている．

本稿では，自社や各企業がどのようなカテゴリの製品を開発しているのかを俯瞰する機能をもつ企業情報比較支援システムについて述べる．

2．企業情報比較支援システム

企業情報比較支援システムは，縦軸を企業名，横軸を製品カテゴリとする表形式によって，個々の企業がどのような製品カテゴリを掲載しているかを表示し，ユーザが個々の企業の製品戦略を比較できるようにする．

そのためには，企業サイトから製品カテゴリに関する情報を収集する必要があるが，製品カテゴリや製品名をあらかじめすべて辞書登録しておくことは困難である．そこで，各サイトのサイトマップを利用し，そこから製品カテゴリや製品名を抽出する．すなわち，サイトマップでは，図1(a)のように製品カテゴリまたは製品名がリンクになっているタイプや，図1(b)のように製品情報全般のページへのリンクになっているタイプが多いと考えられる．

(a)の場合には，リンクのアンカー文字列に製品カテゴリや製品名が網羅して記載されている．したがって，サイトマップのどの範囲にそれらの情報が記載されているのかを決定すれば，アンカー文字列を抽出することで，製品カテゴリや製品名を得ることができる．

一方，(b)の場合には，製品情報を示すアンカー文字列からリンクをたどり，製品カテゴリに関する情報が見つかるまで，順次探索すれば，製品カテゴリや製品名を網羅して得ることができる．

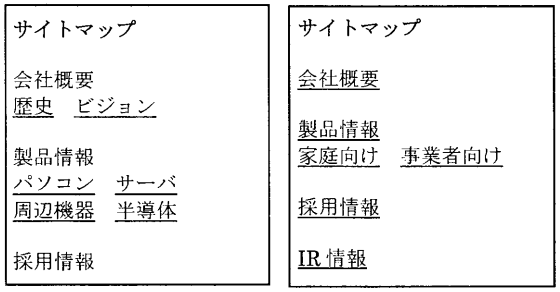
我々は，このようなサイトマップの特徴を利用して，製品カテゴリや製品名を抽出する．

サイトマップから製品カテゴリの記載範囲を限定し，さらにリンク先を探索するかどうかを決定するために，トップページから得られたリンク情報と製品カテゴリオントロジを利用する．

例えば，トップページにあるリンクのうち，リンク先が製品ページでないと判断したものについて，それがサイトマップにも存在したときに，製品カテゴリの記載範囲の切れ目になると判断する．例えば，「採用情報」という単語がトップページに出現していて，それが製品情報を示す単語でないと前もって判断されていれば，図1(a)において，アンカー「採用情報」が製品カテゴリの範囲の切れ目と判断する．

また，製品カテゴリオントロジは，製品カテゴリ間の上位下位関係を表す．例えば，図1(a)において，アンカー「周辺機器」をたどると，周辺機器の下位カテゴリとして，「プリンタ」「スキャナ」等があることが予想される．それは，「周辺機器」の下位概念として「プリンタ」「スキャナ」等があることを人間が知っているからである．そこで，製品カテゴリオントロジを参照して，下位概念があるアンカーに対して，さらにリンクを探索し，リンク先のページから製品カテゴリを抽出する．

さらに，製品カテゴリオントロジは，各カテゴリの同義語を格納し，企業情報を比較する際に，各企業の製品カテゴリ名の多様性を吸収するのに利用される．



(a) 例1　(b) 例2

図1：サイトマップの例

† 沖電気工業株式会社 Oki Electric Industry Co., Ltd.

企業情報比較支援システムは，図２に示す次の構成要素をもつ．

(1) 情報収集部
(2) 文書分類部
(3) 企業情報データベース
(4) 製品カテゴリオントロジ
(5) 企業情報比較部

情報収集部は，各企業サイトのトップページからリンクをたどって，企業情報の比較に必要な情報を収集する．このとき，各サイトのプレスリリース情報，製品情報などに応じて，企業情報データベースに登録する内容が異なるため，文書種別に応じて異なる収集方法を実施する．

文書分類部は，収集中に，トップページからのリンク数やタイトル，リンク元のアンカー文字列，URL などを利用して，プレスリリース情報，製品情報，採用情報などのページ分類をする．

企業情報データベースは，情報収集部で収集した企業情報を企業ごとに格納する．

製品カテゴリオントロジは，製品カテゴリ間の上位下位関係などの関係を表す．

企業情報比較部は，ユーザからの表示要求に応じて，企業情報データベースからデータを検索しユーザに表示する．

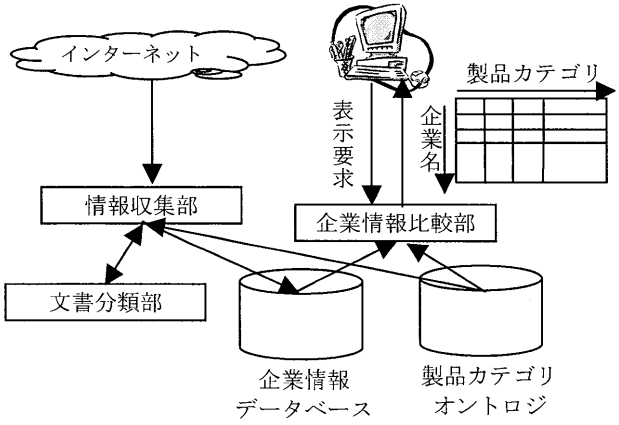


図２：システム構成

3. 情報収集部

3.1 情報収集内容

情報収集部は，各サイトに対して，次の情報を収集する．

[トップページ情報]

トップページのすべてのリンク情報を格納する．リンクのアンカー文字列とリンク先の URL を格納する．例えば，トップページ情報を，表１のように格納する．

[製品カテゴリ情報]

製品カテゴリまたは製品名と，その情報を記載したページの URL を格納する．例えば，製品情報を，表２のように格納する．製品カテゴリオントロジにすべての製品カテゴリをあらかじめ登録することは困難なので，サイトマップのアンカー文字列だけでは，製品カテゴリと製品名を区別できない．したがって，これらの情報の区別はしていない．

[プレスリリース情報]

リリース日とタイトルとそのプレスリリースの内容へのリンクを格納する．例えば，プレスリリース情報を，表３のように格納する．リリース日の順で一覧表示できるように，日付表現の多様性を解消する．なお，本稿では，プレスリリース情報の収集方法は述べない．

表１：トップページ情報テーブル

アンカー文字列	URL	文書種別
会社概要	http://www.**.co.jp/company/	
製品情報	http://www.**.co.jp/products/	製品情報
採用情報	http://www.**.co.jp/recruit/	採用情報

表２：製品カテゴリテーブル

製品カテゴリや製品名	URL
パソコン	http://www.**.co.jp/products/pc/
周辺機器	http://www.**.co.jp/products/peripherals/

表３：プレスリリース情報テーブル

リリース日	タイトル	URL
2004/1/20	新商品が..	http://www.**.co.jp/press/200401
2004/3/28	平成 15 年..	http://www.**.co.jp/press/200403

3.2 製品カテゴリオントロジ

製品カテゴリオントロジは，製品カテゴリ間の上位下位関係や部品関係を表す．図３に製品カテゴリオントロジの例を示す．各オブジェクトが製品カテゴリを表す．製品カテゴリは，様々な表現で文書中に出現するため，それらの多様性を吸収するためにその表現文字列を格納する．例えば，カテゴリ「デジカメ」は，「デジタルカメラ」「デジカメ」などの表現文字列を格納する．

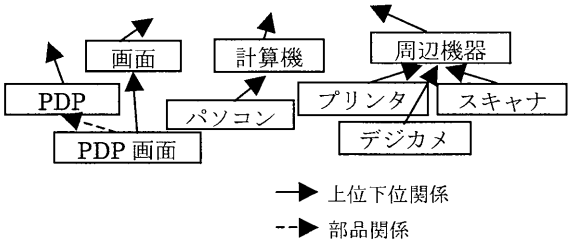


図３：製品カテゴリオントロジ

3.3 製品カテゴリ情報収集方法

次のように製品カテゴリ情報を収集する．

[Step.1] トップページから直接探索できるすべてのページを収集する．収集内容をトップページ情報テーブルに登録する．

[Step.2] Step.1 で収集したページに対して，文書分類を実施する．文書分類では，そのページのタイトル，トップページからそのページへのアンカー文字列，URL を利用する．具体的な決定方法は，松平ら[4]に記載されたヒューリスティックルールを利用する．分類する文書種別は，サイトマップ，製品情報ページ，プレスリリースページ，採用

情報ページ，いずれでもない，のいずれかとして，決定内容をトップページ情報テーブルに登録する．

[Step.3] 製品カテゴリ情報を抽出するため，Step.2 でサイトマップと判断されたページを選択する．そして，図１(a)のように製品カテゴリまたは製品名がリンクになっている範囲の開始位置と終了位置を特定する．サイトマップの文書中に出現する文字列すべてに対して，すなわち，アンカー文字列に限定せず，次のように開始位置と終了位置を決定する．アンカー文字列に限定しないのは，図１(a)の「製品情報」などアンカー文字列以外の単語も，範囲を決定するのに重要な手がかりになるからである．

[開始位置] 文字列が，次の条件１，２のいずれかに合う場合に，その文字列の位置を開始位置にする．

[条件１] その文字列が，「製品」「商品」「ソリューション」「ラインナップ」など製品情報の見出し語になりやすい単語を含む．

[条件２] その文字列が，製品カテゴリオントロジの表現文字列を含む．

[終了位置] 次の条件３，４，５のいずれかに合う場合に，その文字列の前を終了位置にする．

[条件３] その文字列が，「企業情報」「IR 情報」「著作権」など，他に見出し語になりやすい単語を含む．

[条件４] Step.2 で収集したトップページにあるリンクのうち，リンク先が製品情報ページでないと判断したものについて，そのリンクのアンカー文字列を含んでいる．

[条件５] Step.2 で収集したトップページにあるリンクのうち，リンク先が製品情報ページでないと判断したものについて，そのURL が一致する．

例えば，トップページ情報テーブルが表１であると仮定する．この場合，製品情報と判断されなかったのは，第３レコードである．したがって，条件４により「採用情報」を含む文字列が，製品情報を表す領域の終了位置になる．また，条件５により，リンク先が「http://www.**.co.jp/recruit/」であるリンクが，製品情報を表す領域の終了位置になる．図４に，図１(a)の文書に対する領域判定処理を示す．

[Step.4] Step.3 で登録されたリンクのアンカー文字列に，製品カテゴリオントロジの表現文字列が含まれる場合には，リンク先の文書に，表現文字列が一致した製品カテゴリの下位概念の製品カテゴリを含む可能性が高い．そこで，その場合には，リンク先の文書を対象にして，Step.3 の処理を繰り返す．

例えば，図１(a)の文書では，「周辺機器」が一致する．したがって，このリンク先に文書を探索して，Step.3 の処理を繰り返す．

4. 企業情報比較部

企業情報比較部は，縦軸を企業名，横軸を製品カテゴリ名とする表形式によって製品情報を表示する．

横軸は，製品カテゴリオントロジにある製品カテゴリだけでなく，ユーザが自由に設定することができる．

システムは，横軸の個々の製品カテゴリ名をキーワードにして，製品カテゴリテーブルの「製品カテゴリや製品名」項目を検索する．この際，横軸の製品カテゴリ名が，製品カテゴリオントロジに登録されている場合には，対応する製品カテゴリに設定されている表現文字列を用いて検索することで，「製品カテゴリや製品名」項目の表現の多様性を吸収する．また，対応する製品カテゴリに下位概念がある場合には，下位概念の製品カテゴリでも検索する．

さらに，同時に，同じキーワードでプレスリリース情報テーブルの「タイトル」項目を検索することで，関連するプレスリリースも見つけることができる．

カテゴリ比較の出力結果を図５に示す．

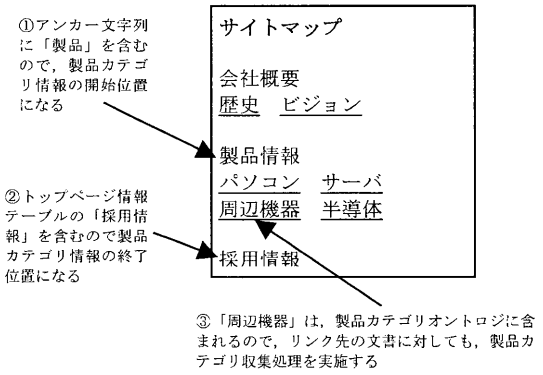


図４：図１(a)の文書に対する処理

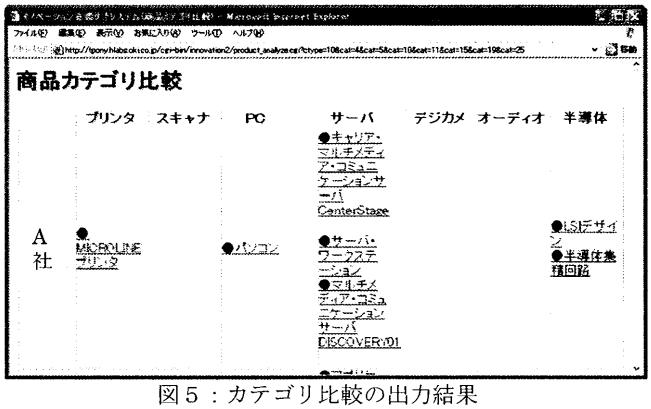


図５：カテゴリ比較の出力結果

5. 実験と考察

企業サイトのうち，サイトマップがどの程度存在し，そのうち，サイトマップをどの程度見つけることができるのか，また，サイトマップから，どの程度の製品カテゴリを抽出できるのかを実験した．

5.1 実験結果

5.1.1 文書種別の判定

１２１個の企業サイトについて，トップページからリンクされているページを抽出し，個々のページについて，文書種別がサイトマップ，プレスリリース，製品情報のいずれになるかを分類した．

表４に分類結果を示す．１２１サイトのうち，目視でトップページからサイトマップ，プレスリリース，製品情報のページへのリンクを含むページが見つかったのは，それぞれ６７サイト，６３サイト，１００サイトあった．そのうち，それぞれのページへのリンクをシステムが抽出できたのは，５４サイト，４４サイト，５４サイトであった．

また，サイトマップまたは製品情報へのいずれかのリンクを含むサイトは１１０サイトあり，そのうち，サイトマ

ップまたは製品情報のいずれかを正しく抽出できたのは，６９サイトであった．

表４：文書分類結果

文書種別	適合率	再現率
サイトマップ	100.0% (54/54 サイト)	80.6% (54/67 サイト)
プレスリリースページ	88.0% (44/50 サイト)	69.8% (44/63 サイト)
製品情報ページ	80.6% (54/67 サイト)	54.0% (54/100 サイト)
サイトマップまたは製品情報ページ	84.2% (69/82 サイト)	62.7% (69/110 サイト)

5.1.2 製品カテゴリ名と製品名の抽出

5.1 節でサイトマップと判断された５４サイトのうち，製品カテゴリや製品名を含まない，不動産系やアミューズメント系の企業を除いた４２サイトを対象として，サイトマップから製品カテゴリ名または製品名を示すリンクを抽出した．但し，3.3 節の Step.3 までの処理で評価を行った．製品カテゴリオントロジとして，１０数語の情報関係の用語を利用した．

表５に抽出結果を示す．４２のサイトマップのうち，製品カテゴリ名または製品名を示すリンクは，１０５７個存在した．したがって，サイトマップ中の製品カテゴリ名または製品名へのリンク数は，平均２５．２個になる．

そのうち，８０８個を製品カテゴリ名または製品名を示すリンクとして正しく抽出した．図６に抽出された製品カテゴリを示す文字列の例を示す．抽出された文字列には，製品名を含んでいるものや，複数の製品カテゴリ名が，「／」や「・」で区切られているものが見られる．

表５：製品カテゴリ抽出結果

正解数	適合率	再現率
1057	70.4% (808/1147 個)	76.4% (808/1057 個)

5.2 考察

１２１サイトのうち，サイトマップが存在したのは，６７サイトであり，サイトマップだけでは，製品カテゴリや製品ページを見つけるのは難しいことがわかる．しかし，トップページから，直接製品情報ページへのリンクがある企業を含めると，１１０サイトあることから，これら２種類の文書情報を組み合わせる必要がある．

表４において，サイトマップに関しては，適合率，再現率とも比較的高かったが，製品情報ページに関しては，再現率が特に低かった．これは，トップページに製品情報ページへのリンクが存在する場合には，製品カテゴリ名がアンカー文字列になっている場合が多く，製品情報ページと判断できなかったからである．今後，文書種別の分類にオントロジ辞書を使う必要がある．

すべての文書種別に共通して再現率を低下させた理由として，Java スクリプト等を用いてリンクが記載されている場合に，リンクを抽出できない点が挙げられる．

表５においては，オントロジ辞書の登録数が十分でなかったにも関わらず，良好な結果が得られた．

しかし，再現率はもっと良いと予想していた．それは，本手法では，サイトマップから製品カテゴリや製品名を示す領域の終了位置を決めるのは困難であるが，開始位置を決めるのは容易だと考えていたからである．しかし，開始位置がうまく決まらないことが多かった．これは，開始位置を示すキーワードとして，「製品」などのアンカー文字列が画像ファイルになっていること，製品カテゴリオントロジの登録数が少なく十分機能しなかったことが挙げられる．

本稿では，3.3 節の Step.4 の実験結果を報告していないが，今後実験を行う予定である．

IP 多機能電話機 e 音 IP フォン
IP ネットワーク製品（ルータ・LAN スイッチ）
構内 ADSL
モデム・TA
電話機・交換機
半導体集積回路
光コンポーネンツ
ATM・自動契約機
フィルムカメラ／双眼鏡
デジタルカメラ
デジタルビデオカメラ
液晶プロジェクター
複合機／コピー機／ファクス
通信／TV 会議／ソフトウエア
イメージ・デジタルフォトシステム
粉体塗装機器
プレートコイル

図６：製品カテゴリを示す文字列の例

6. おわりに

我々は，新商品の企画全般を支援するシステムを開発することを目標として，自社や各企業がどのようなカテゴリの製品を開発しているのかを俯瞰する機能をもつ企業情報比較支援システムについて述べた．

今後の課題は次の通りである．

- アンカー文字列の製品カテゴリ名と製品名を判別する方法の検討．
- トップページやサイトマップから製品情報ページへのリンクをさらにたどって，リンク先の製品カテゴリ名と製品名を抽出する実験の実施．

参考文献

[1] 嶋田和孝,遠藤勉:“製品性能表からの特徴データの抽出”,情報処理学会 NL 研報,133-15, pp.107-113,1999.
[2] 徳永秀和,天雲勇作,青江順一:“ソフトウエア製品ニュースからの開発傾向の抽出”,情報処理学会 NL 研報,159-3, pp.13-18,2004.
[3] 松平正樹,上田俊夫,大沼宏行,森田幸伯:“Web コンテンツの分析に基づくオントロジー構築および情報整理の試み”人工知能学会,SIG-SWO-A302-08,2003.