7V-4

商品情報サイトの横断検索と商品情報の比較を支援するシステムの開発と評価

米坂 元宏† 植木 泰博‡ 堀川和義* 冬木 正彦§ 荒川 雅裕§

関西大学大学院工学研究科† 関西大学先端科学技術推進機構‡ 株式会社e-kikai* 関西大学工学部§

1. はじめに

インターネット上には多くの情報を提供するページが存在し，情報は散在している．異なるサイト上の同種の情報を比較することは容易ではない．容易に閲覧・比較するためには，情報を効率よく抽出し統合する必要がある．この問題解決のために，検索エンジンが用いられている．

検索エンジンの技術発展は，福島[1]によると次のよう分類されている．

人手による収集によって構築されたディレクトリ型検索エンジンが第一世代，クローラーによるWEBページの収集を自動化したロボット型検索エンジンが第二世代，Google[2]に代表されるリンク解析に基づく新しい検索結果ランキング方式を導入した第三世代．近年，検索エンジンの利用人口増加に伴い，利用目的が多様化している．新たな発展として，目的特化型の検索エンジンが第四世代に位置づけられている．

第三世代の検索エンジンでは，対象とするWEBページの規模の増加に伴い，検索できる情報の鮮度を維持することが難しくなる．一方，目的特化型検索エンジンは，目的を限定することで検索対象とすべきウェブページの規模を抑え，巡回サイクルを短縮して情報鮮度を向上させることが可能になる．したがって，目的に応じた，より詳細な検索・分類機能を提供することが可能になる．

目的特化型の検索エンジンの要素技術には，フォーカストクローラや情報抽出など，ウェブページの収集・登録時の要素技術に特徴がある．

日本における目的特化型の検索エンジンの事例として，商品の価格比較を行うYahoo!商品検索[3]，求人情報に特化したジョブエンジン[4]，ブログの情報に特化したblogWatcher[5]などがある．

本研究では，商品情報の閲覧や比較など，手間と時間のかかる作業の負担を軽減させることを目的とし，中古機械の商品情報を扱う複数のWEBサイトを対象とした横断検索エンジンと商品情報の比較を支援するWEBアプリケーションのシステム開発を行う．さらに，システムの運用テストを行い，実用性を評価する．

Development and Evaluation of A System for Merchandise Information Sites Cross-Search and Merchandise Information Comparison Support

† Motohiro Yonesaka, Kansai University

‡ Yasuhiro Ueki, ORDIST, Kansai University

* Kazuyoshi Horikawa, e-kikai Corporation

§ Masahiko Fuyuki, Kansai University

§ Masahiro Arakawa, Kansai University

2. システムの分析

2.1 中古機械業界のWEBサイト

目的に沿う検索エンジン構築のために，中古機械業界のWEBサイトの特徴について調査した．

中古機械業者に協力していただき，主要な中古機械取引サイトを22サイト選定した．これらに登録されている機械数は約23,000件ある

これら22サイトの商品情報はすべてHTMLの表形式で記述されている．中でも単純な表形式の構造として，一列目が属性で2列目が属性値のセットが複数行ある，もしくは1行目に属性で2行目以降が属性値のセットとなっているサイトが22サイト中16サイトを占める．

各サイトの機械情報は，1日から1週間に1回の間隔で更新される．

2.2 機能要件

上述したWEBサイトの特徴を考慮し，システムの機能要件を検索エンジンの構成要素別に分析した．

2.2.1 クローラー部

質の向上を図るため対象とするドメインを限定し，クローラーはWEBサイトのリンクを限なく巡回する．各WEBページ内に表形式で記述された商品情報を表１の属性に対応するように抽出する．まずは2.1で述べた単純な表形式の構造をもつ16サイトに対応するようシステムを開発する．

情報の鮮度を保つために，最低1日1回は各サイトを巡回し，情報の存在の有無や新規登録された情報を更新する．

2.2.2 インデクサ部

クローラーで収集した商品情報をデータベースに蓄積し，高速な検索処理を可能にするために全文検索エンジンを用いる．専門用語が多いので，N-gram方式による漏れのない検索ができ，属性検索が可能な全文検索エンジンが必要である．

2.2.3 検索部

検索機能の他，情報の比較を支援にするために以下の機能をWEBアプリケーションで実現する．商品情報の特徴による比較表の表示機能，日々の新着情報表示機能，気に入った機械情報の保存機能とメモ機能，保存した機械情報の存在有無の表示機能．

表１　抽出する商品情報の属性

機械名	メーカー	型式	仕様
年式	価格	会社名	保管場所

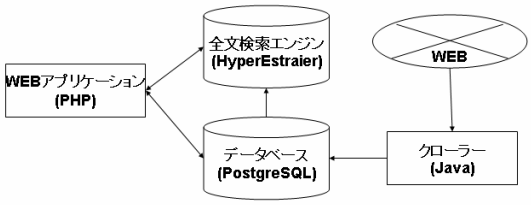

図１　システムの構成

図２　開発した WEB アプリケーションの画面

3. システムの開発

システム構成を図1に示す．

3.1 クローラー部

クローラーは，幅優先探索方式で各WEBサイトの同一ドメイン内のリンクを巡回する．

各WEBページから目的の情報を抽出する方式として代表的なものは，抽出例から機械学習を行う方式（学習ベース），抽出対象の文字列の辞書を用いる方式（辞書ベース），ページ構造の手がかりを用いる方式（ヒューリスティックベース）がある．

各抽出方式では商品についての知識を入力する作業が必要である．この作業にかかる人的コストを小さくせざるを得ない場合で，単純に抽出した値を表としてユーザに提示し，再現率を優先する場合は，ヒューリスティックベースが優れている．[6]

したがって，ヒューリスティックベースを用いて，HTML中の表の中から商品情報を抽出し，データベースに蓄積する．この場合，同一の機械情報でも表示形式が異なっていれば，クローラーは異なる機械情報とみなす．さらに，抽出済みの情報の更新機能も実装した．

3.2 インデクサ部

データベースに蓄積した情報から，全文検索エンジンHyperEstraier[7]を用いて，インデックスを作成する．インデックスは検索部で用いられる．

3.3 検索部

検索部はWEBアプリケーションで作成し，ユーザ認証を設け，ユーザ個別の情報を保存ができるようにした．開発した一画面を図2に示す．

4. システムの運用テストと評価

4.1 クローラーの検証

上述した16サイトの登録機械数を満たす機械情報をクローラーが抽出したことを確認した．

さらに，クローラーが正しく情報を取得できているか詳細に調べるために以下の方法で検証を行った．

まず，毎日クローラーで機械情報を収集しリスト1を作成する．次に，手作業で求めたい機械情報を収集しリスト2を作成し，リスト1とリスト2を照合する．

機械情報の登録数で大中小規模にグループ分けを行い，代表的な3サイトで照合した．リスト2を2005年9月16日時点で作成したところ，登録機械数はそれぞれ9543件，658件，270件であった．

リスト1とリスト2を照合した結果，リスト2の機械情報はリスト1にすべて含まれていることを確認した．

4.2 ユーザによる評価

システムの実用性を検証するために，評価用サイトを設置し，日常業務で機械情報の検索を行っている業者3社に協力していただき，3人のユーザに1週間～2週間程利用していただき，後日アンケート及びヒアリングを行い，以下のような意見を得た．

機能性について，「多数のサイトごとのチェックをしなくて済む」「即日，又は翌日くらいには情報が入るので非常によいと思う．情報量が多いので良い」「市場にどういう種類の機械が出てきたかがつかめる」「保存しておいた機械の情報を比較することで，他社の売れ行きの機械の情報を得ることができる」という意見を得た．したがって，情報の鮮度を保ちつつ，複数サイトの機械情報を比較する負担を軽減するニーズを満たしていると考えられる．

作業時間の変化について，「あまり変わらない」「時間的には同じ位だが，見逃しが少なくなったと思う．同じ時間で多くの機械を探せるようになった」という意見を得た．したがって，同じ時間で多くの機械を探せるようになるなど，検索の質が向上したと考えられる．

以上の結果から，情報の鮮度を向上させ，複数のサイトの商品情報の特徴比較のために情報を効率よく抽出し統合してユーザに提供する実用的な検索エンジンを実現できた．

参考文献

[1] 福島俊一:「検索エンジンの仕組みと技術の発展」情報の科学と技術 (54巻2号, 2004, pp.66-71.)
[2] Google：http://www.google.com
[3] http://psearch.yahoo.co.jp/
[4] http://www.jobengine.jp/
[5] http://blogwatcher.pi.titech.ac.jp/
[6] 楠村幸貴，土方嘉徳，西田正吾：「e-マーケットプレース向け情報収集・抽出方式における人的コストの調査」情報処理学会研究報告 (Vol.2005,No.42(DBS-136 FI-79),P63-70)
[7] http://hyperestraier.sourceforge.net/