D-039 検索サイトにおける入力項目と検索結果のフィールド名の対応調査(D.データベース)

Ohmori Keisuke
九州大学大学院システム情報科学府

Nakatoh Tetsuya
九州大学情報基盤センター

Hara Yukari
九州大学情報基盤センター

Hirokawa Sachio
九州大学情報基盤センター

出版情報：情報科学技術フォーラム一般講演論文集. 3 (2), pp.89-90, 2004-08-20. FIT(電子情報通信学会・情報処理学会)推進委員会
バージョン：
権利関係：

D-039

検索サイトにおける入力項目と検索結果のフィールド名の対応調査

大森 敬介 † **中藤 哲也** ‡ **原 由加里** ‡ **廣川 佐千男** ‡
Keisuke Ohmori Tetsuya Nakatoh Yukari Hara Sachio Hirokawa

1. はじめに

WWW 上で提供されている情報検索のサイト (以下，**検索サイト**と呼ぶ) には，Google などの一般検索エンジンのサイトの他に，自サイト内のデータベースに対する検索を提供するサイトも数多く存在する．それらのデータベースの情報は，検索によってデータベースから動的に生成されるページによってのみ参照されるため，直接参照する事はできない．そのため，これらは Invisible Web, Deep Web あるいは Hidden Web などと呼ばれる．

これらの検索サイトは特定のテーマに限定した質の高い情報を提供している事が多い．我々は，そのような複数の検索サイトに対してメタサーチを行うシステム **DAISEn** [5] を開発している．一般的なメタサーチエンジンは統合対象である検索エンジンが固定であるが，本システムは目的の検索エンジンを自動的に分析しメタサーチエンジンを動的に生成する．

近年検索サイトには，これまでとは異なる新しい方向性がみられるようになって来た．それらの検索サイトにおいては，複数の入力フィールドを用いた検索に対して，入力項目に一致するレコードのリストが返される．例えば，Amazon.com は本の情報に関するレコードのリストを返す．kakaku.com は PC のリストと共にそれらの価格を返す．Travelocity.com は指定されたエリアのホテルに関するレコードのリストを返す．

WWW 上のこのようなデータベースを横断的，統合的に検索を行なうためには，各データベースが扱うレコードが何であるかが予め分からなければならない．従来のデータベースや WebService であれば，データスキーマも明示的に与えられている．しかし，検索サイトはブラウザ経由の利用しか想定されていない．従って，各検索サイトが扱うデータスキーマは，入力ページのフォーム情報や検索結果出力情報から抽出，推定しなければならない [3].

我々は，メタサーチシステム DAISEn をこれらの検索サイトに対応させるため，複数の入力項目を持つ検索サイトを収集，分析している．我々は，国立国会図書館のデータベースナビゲーションサービス **Dnavi** [1] に登録されている検索サイトについて詳細な調査を行なっている [4, 2]．またフィールド名の自動抽出に関する研究 [1] を行なっている．

本論文では，実際の検索サービスを対象に入力項目と検索結果のフィールド名の対応を調査した．部分的にでも入力項目と検索結果のデータスキーマを抽出し対応をつける事で，より正確なデータスキーマが構成できると考えられる．

2. 複数の入力項目を持つ検索サイト

単純なテキスト検索を行なう検索サイトの多くは，一つのテキスト入力フィールドを持ち，キーワードを入力する事によってそのキーワードを持つテキストへのリンクを補助情報と共に提示する．

一方，複数の入力項目を持つ検索サイトの多くは，データベースの持つフィールドの一部を入力する事でそれに一致するレコードのリストを出力する．そのようなサイトの実例を図 1 に示す．この書籍検索サイトでは，書籍に関する複数のフィールドを指定する事で，一致する書籍データの一覧をユーザに提示する（図 2）．検索結果のフィールド名が入力項目のフィールド名と対応している事が分かる．

図 1: 複数の入力項目を持つ検索サイトの例

図 2: 検索結果の例

†九州大学大学院システム情報科学府
‡九州大学情報基盤センター 〒 812–8581 福岡市東区箱崎 6–10–1
[1] http://dnavi.ndl.go.jp/

3. フィールド名の調査

我々は Dnavi 中に登録されている 2,880 件の検索サイトを詳細に調査している [4]．この 2,880 件の検索サイトのうち複数のテキスト入力フィールドを持った検索サイト 1,541 件から無作為に 100 件の検索サイトを選び，今回の調査の対象とした．

検索における入力項目及び検索結果のフィールド名抽出は次の方法で行った．

入力項目におけるフィールド名は，テキスト入力フィールド[2]に関連付けされていると思われる文字列を人手で抽出した．関連付けされているのが文字列ではなくプルダウンメニュー[3]である場合は，そのメニューの内容を全てを該当入力項目におけるフィールド名とした．

検索結果におけるフィールド名も，結果の各フィールドに関連付けされていると思われる文字列を人手で抽出した．検索結果がリンクのリストで示されておりフィールドを持たないタイプの場合は，リンク先のページを調べ，同様の方法でフィールド名を抽出した．

入力項目と検索結果の共通フィールドの割合 R を次式で求めた．

$$R = \frac{\text{入力項目と検索結果に共通のフィールド数}}{\text{入力項目のフィールド数}}$$

4. 結果と考察

調査結果を図 3 のヒストグラムに示す．35% のサイトでは，入力項目のフィールドが全て検索結果のフィールドに現れている．それらは入力した一部のフィールド値から，同じフィールド値を持つレコードの全フィールドを得るタイプの検索サイトであると言える．また，入力項目の半分以上のフィールドが出力フィールドに現れているのは，全体の 84% である．

この図の中で共通フィールドの割合が 0〜0.1 の範囲としてカウントされている 13 サイトの実際の値は全て 0 である．これらのうち 11 サイトでは，入力項目のフィールド名が “キーワード” かそれに類するものであり，特定の項目名は指定されていない．しかし，出力結果において名前が付いた項目中にこのキーワードが出現する事が多い．従って，入力において単なる “キーワード” ではなく，具体的な項目名を付けることも可能と考える．

共通フィールドの割合が 1 もしくは 0 以外の約半数のサイトについては，割合は 0.4〜0.9 の範囲に集まっている．これらのサイトの多くでは，入力項目のフィールド名のうち検索結果に現れないものとして “キーワード”, “内容”, “ISBN” や “ISNN” が頻出する．“キーワード” に関しては前述の通りである．“内容” は “キーワード” と同じように特定のフィールドを対象としない場合，及び拡張項目としてレコード毎にその有無が異なる場合があった．“ISBN” や “ISNN” に関しては，多くの検索サイトにおいて “図書番号” のようなあいまいなフィールドとして検索結果に出現していた．よってこれらのフィールド名はデータベースが持つものではなく，入力のために便宜的に用意されたフィールド名と思われる．

[2] `<input type=text>` で生成される要素
[3] `<select>` で生成される要素

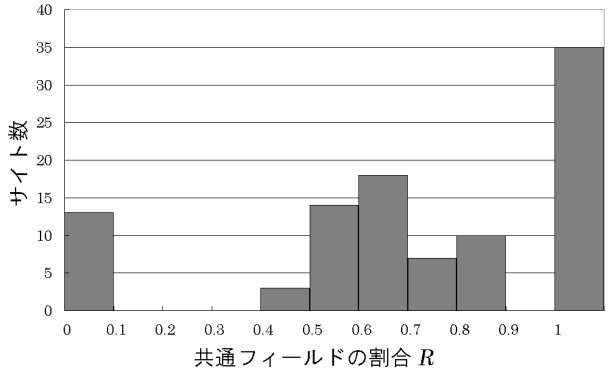

図 3: 共通フィールドの割合 R のヒストグラム

今回の調査には現れなかったが，出力項目のフィールド名が省略されている検索サイトについても共通フィールドの割合が 0 となる．そのようなサイトはデータスキーマを入力フィールドの情報から抽出する事が必要と考えられる．

5. まとめ

無作為に抽出された検索サイトを対象に，入力項目と結果出力のフィールド名の対応を調査し，それらが強い相関を持つ事を明らかにした．これはデータスキーマの抽出において，それらのフィールド名が有用な情報になる事を示す．

今後更に詳細な調査を行なうと同時に，他の手法による推定も検討し，それらの組み合わせによりデータスキーマの自動抽出システムを完成させる予定である．

参考文献

[1] 大森敬介, 中藤哲也, 山田泰寛, 原由加里, 廣川佐千男， 複雑な検索機能を持つ検索サイトの動向調査 DEWS2004, I-1-05, 2004.

[2] T. Nakatoh, K. Ohmori, Y. Yamada and S. Hirokawa, *COMPLEX QUERY AND METADATA*, Proc. ISEE2003, pp. 291-294, 2003.

[3] S. Thakkar, C. A. Knoblock, J. Ambite and C. Shahabi, Dynamically Composing Web Services from On-line Sources, Proc. of 2002 AAAI Workshop on Intelligent Service Integration, Edmonton, Alberta, Canada.

[4] 山田泰寛, 松永吉広，野口正人，中藤哲也，廣川佐千男， 統合検索システム DAISEn での検索サイトフォーム分析， 情報処理学会研究報告 2003-DBS-131(II)(77) (DBWS2003), pp.311-318, 2003.

[5] 専門検索サイトの動的統合による次世代検索システム DAISEn， Directory Architecture for Integrated Search Engines, `http://daisen.cc.kyushu-u.ac.jp/`