D-5

検索サイトのための集合演算子の自動推定

Automatic Estimation of Set Operator for Search Sites

中藤 哲也† 酒井 美由紀‡ 廣川 佐千男†
Tetsuya Nakatoh Miyuki Sakai Sachio Hirokawa

1．はじめに

膨大なWWWの世界から必要な情報を効率よく得る為，我々はYahoo!やgoogleなどのようなWWW全体を対象とする検索エンジンを利用し，閲覧する範囲を絞り込む．しかしながら，求める情報とは無関係なページが多く含まれることも多く，検索結果の質が問題となっている．

一方，企業などのWebサイトにおいては，自サイト内の情報やデータベースについての検索機能を用意しているところが増えている．本稿ではこのようなサイトを検索サイトと呼ぶ．検索サイトは，そのサイト内の情報を効率良く提供することを目的としており，一般の検索エンジンより高品質の情報が期待できる．

しかしながら，検索サイト自身が持つ情報量は一般の検索エンジンに比べると少ない傾向にあり，利用者が必要な情報を集めるためには多くの検索サイトを調べる事が必要である．そのため，そのような複数ある検索サイトの一つ一つに対して個別に検索を行なわなければならない．

我々は，このような検索サイトに対して一括して検索を行ない，結果を一覧にまとめてユーザに提示する検索統合システムを開発してきた [7]．検索サイトに対するクエリーの自動生成 [5]と結果抽出を行なうラッパーの自動生成 [2]，自動生成したラッパーの精度推定のための検索結果一覧における表示件数の推定手法の提案 [3]を行なっている．ラッパーを使った適切な特徴ベクトル生成 [1]とそれを使った検索サイトの分類 [4]の試みも行っている．

検索サイトの多くは，単一のキーワードを指定する事による単純な検索に加えて，複数のキーワードを指定して何らかのクエリーを構成する事による検索が可能である．しかしながら，そのクエリーの構成方法(演算子の使い方)は，各検索サイトの運営方針や使用している検索エンジンによって異なり，統一性はない．

検索サイトに対する統合検索において，複数のキーワードを用いた検索を行なうためには，検索サイト毎に異なるインターフェースをユーザに対して隠蔽し，統一された操作方法を提供するのと同様に，それらクエリーの違いも隠蔽する事が必要である．

我々は，このうちクエリーの自動推定について既に一部明らかにしている [6]が，本稿では，差集合演算を含めた集合演算子の自動推定について，その手法の提案を行なう．

2．検索用集合演算子の自動推定

単一のキーワードを使用した単純な検索は，全ての検索サイトで可能であるが，それに加えて多くの検索サイトでは複数のキーワードを使った検索が可能である．複数のキーワードをそのまま並べて，あるいは何らかの演算子を用いて指定する事で，検索範囲を広げたり狭めたりする事が出来，より的確に必要とする情報を選び出すことが可能となる．

それらの集合演算子の使用方法は多くの場合，各サイトの検索フォームが存在するWebページ自体やそこからリンクによって辿ることの出来るWebページに自然言語によって記載されている．

非常に多くの検索サイトを扱うためには，この検索用の集合演算子も自動的に推定する必要がある．Webページ上に記載された説明文を解析すれば必要な情報は得られるが，それには自然言語解析(理解)が必要であり容易ではない．

我々は，想定される演算子を構成し，それらを使用して実際に検索サイトにおいて検索を行ない，その結果の解析によって，検索用の集合演算子を推定する方法を提案する．この方法により，サイト毎に必要なクエリーを的確に生成でき，またヒューリスティックも必要ない．

2.1 集合演算子の推定方法

検索サイトに対してキーワードを与えて行なう検索とは，そのサイトの持つ全文書(情報)の集合の中からキーワードを含む(関連する)文書(情報)からなる部分集合を取り出す操作に他ならない．複数のキーワードを与えて検索を行なう場合は，各キーワードによって得られる部分集合に対してどのような集合演算を行なうかによって，得られる結果が異なる．

集合に対する基本的な演算としては，和集合(union)，共通集合(intersection)，差集合(set difference)がある．実際の検索サイトにおいても，これら３つの演算を行なう為の演算子が用意されている事が多いので，それらを推定対象とした．

一般的に用いられる演算子を候補として仮のクエリーを構成し，そのクエリーで実際に検索を行なって得られた結果の件数より，その候補がクエリーとして有効な記述であったかを判定する．検索結果から結果の件数を求める技術は既に有している．[2][3]

演算子の候補には次のようなものが考えられる．

- 和集合を求める演算子　　**OR or + | ,**
- 共通集合を求める演算子　　**AND and & ***
- 差集合を求める演算子　　**NOT not - ! #**

この判定を的確に行なうためには，次に示すような２種類に分類されるキーワードが必要である．

A: 各検索サイトにおいて検索結果が得られる
Z: 各検索サイトにおいて検索結果が０件である

対象となる各検索サイトにおいて，候補のキーワードで事前に検索を行ない，その検索結果を判定することで，この２種のキーワードを正確に選択できる．

候補の各演算子(***op***)にこの２種のキーワードを組み合わせる事で得られた検索結果は，基本的には表１となる．しかし，候補として検索をテストした演算子が機能しなかっ

†九州大学 情報基盤センター
‡九州大学 大学院 システム情報科学府

た場合など，常に表１がそのまま適用出来るわけではない．その点をこれより詳細に見てゆく．

表１：検索結果とそれが意味する集合演算

A op A	*A op Z*	*Z op A*	*Z op Z*	機能
+	+	+	0	和集合
+	0	0	0	共通集合
0	+	0	0	差集合

凡例 +:結果あり 0:結果が０件 ?:不定 (今後の表で共通)

2.2 演算子無しの場合の機能

キーワードを空白のみで区切って並べる事でクエリーを構成した場合，一般的には和集合か共通集合が得られる事が多い．候補である演算子を用いた検索を行なう前に，この演算子を使わない場合にクエリーがどのような意味を持つかを推定する必要がある．何故ならば，候補の演算子が演算子として働かずに無視された場合，キーワードの一つとして見なされ，空白のみで区切ったクエリーとなる為である．

複数のキーワードを空白のみで区切った場合に予想される結果を表２に示す．

表２：演算子無しの場合

A A	*A Z*	*Z A*	*Z Z*	機能
+	+	+	0	和集合
+	0	0	0	共通集合
+	+	0	0	2項目以後無視
0	0	0	0	エラー

2.3 演算子無しで共通集合が得られた場合

演算子を使わずにキーワードを空白で区切ることで共通集合が得られる場合，検索結果を表３に当てはめる事で，演算子の機能が推定出来る．

表３：基本が共通集合演算の場合

A op A	*A op Z*	*Z op A*	*Z op Z*	*op* の機能
+	+	+	0	和集合
+	0	0	0	共通集合
0	+	0	0	差集合
?	0	0	0	演算子無機能
0	0	0	0	エラー

表３から，候補の演算子が何の機能も持たなかった場合と，共通集合の演算子であった場合の判別が付かない事が分かるが，演算子無しで共通集合が得られるので，単純に，和集合と差集合の演算子のみ求めればよい．

2.4 演算子無しで和集合が得られた場合

演算子を使わずにキーワードを空白で区切ることで和集合が得られる場合，検索結果を表４に当てはめる事で，演算子の機能が推定出来る．

表４：基本が和集合演算の場合

A op A	*A op Z*	*Z op A*	*Z op Z*	*op* の機能
+	+	+	0	和集合
+	0	0	0	共通集合
0	+	0	0	差集合
+	+	+	?	演算子無機能
0	0	0	0	エラー

表４から，候補の演算子が何の機能も持たなかった場合と，和集合の演算子であった場合の判別が付かない事が分かるが，演算子無しで和集合が得られるので，単純に，共通集合と差集合の演算子のみ求めればよい．

2.5 演算子が必須であった場合

演算子を使わずにキーワードを空白で区切っただけでは検索が出来なかった場合，検索結果を表５に当てはめる事で，演算子の機能が推定出来る．

表５：演算子が必須であった場合

A op A	*A op Z*	*Z op A*	*Z op Z*	*op* の機能
+	+	+	0	和集合
+	0	0	0	共通集合
0	+	0	0	差集合
+	+	0	0	演算子無機能
0	0	0	0	エラー

3. 実験と評価

上記の推定方法を，9 つの実際の検索サイトに適用し，集合演算子の推定を行った．

得られた演算子表現を，各検索サイトにあるヘルプ等の記載と照らし合わせて，一致しているかを調べたところ，次のような結果が得られた．

基本的に一致：6サイト

ヘルプの内容以外の表現も発見：2サイト

ヘルプの内容と不一致：1サイト

不一致だった１サイトについて，手動で確認したところ，我々の推定に従って正しい検索結果が得られる事，ヘルプに記載の方法では正しい検索が出来ない事を確認した．このようにヘルプの記載が間違っている場合もあり，ヘルプの記載を信用するよりもむしろ我々の手法により正しい演算子が得られる事が明らかになった．

4. まとめ

検索サイトの統合に必要な技術として，複数のキーワードで検索を行なう時に使用する集合演算子を自動的に推定する手法について提案した．今後，より多くの検索サイトでテストし，検索統合システムに実装する．

参考文献

1) S. Hirokawa, S. Watanabe, Y. Koga, T. Taguchi: Automatic Feature extraction of Search sites, Proc. SSGRR2001(CD-ROM) (2001).
2) 古賀 康則，田口 剛史，廣川 佐千男: 検索サイト統合のためのラッパー生成法, DEWS2001 CD-ROM:6b-1 (2001).
3) 古賀 康則，酒井 美由紀，廣川 佐千男: 統合検索のための検索結果件数推定方法, 第 63 回情報処理学会全国大会（2001）.
4) 中藤 哲也，古賀 康則，廣川 佐千男: 検索統合のための検索サイト分類法, Proc.DBWeb2001 pp.225-228(2001).
5) T. Nakatoh, M. Sakai, Y. Koga, S. Hirokawa: Generation of Query URL for Search Sites, Proc. SSGRR2002w(CD-ROM) (2002).
6) T. Nakatoh, Y. Koga, A. Uhl, S. Hirokawa: Automatic Estimation of Query Form for Search Sites, Proc. PYIWIT'02 (2002).
7) T. Taguchi, Y. Koga, S. Hirokawa: Integration of Search Sites of the World Wide Web, Proc. Intern. Forum cum Conf. on Information Technology and Communication, Vol.2, pp25-32 (2000).