2U－8

情報フィルタリングの手法を用いた転送先学習型メタ検索エンジン

加藤大志 沼尾正行
東京工業大学計算工学専攻

1 はじめに

近年、インターネットの技術は急速に普及し、コンピュータの世界を一新するほどの力を持つようになってきた。インターネットの技術で最も広く知られているものは World Wide Web(以下 Web) である。Web の急速な拡大により、世界中に存在するページは数えきれぬほど多量になった。ユーザの欲しい情報は Web のどこかに存在すると言っても過言ではないが、Web の広大さゆえにその情報を見付けるのは非常に困難である。

その困難さを解決しようと登場したのが検索エンジンである。検索エンジンは、内部にデータを持ちユーザからの検索式をデータに照らし合わせ結果を出力するロボット収集型とディレクトリ登録型、及び、それらの検索エンジンに検索式を転送するメタ型に大別される。ところが、検索エンジンの数が増加するにつれ、ユーザがどの検索エンジンを使えばよいのか選択することも困難になってきた。これに対処するための一つの方法として、検索エンジンの自動選択が考えられる。

本稿では、メタ検索エンジンにおいて転送先を検索式に応じて順位付けする情報フィルタリングを提案し、この手法を実装した転送先学習型メタ検索エンジン MetaRoamer について述べる。

2 SemiCOLON:述語論理フィルタリングシステム

MetaRoamer では検索式から検索エンジンの順序を推論するために、述語論理フィルタリングシステム SemiCOLON を使用している。

SemiCOLON は述語論理の非決定性をコストを用いて制御するシステムである。SemiCOLON に与える知識 (節) には数値のコストが付けられ、推論時には推論経路の各節のコストの平均が小さいものから推論される。また、推論による節の呼び出しでコスト減少、バックトラックでコスト増加とすることで節のコストを変化させ学習効果を期待する。

例えば、SemiCOLON に与えられた知識が、

コスト	節
20001	`fruit(orange).`
20002	`fruit(apple).`
20001	`food(X):-fruit(X).`
20000	`food(X):-vegetables(X).`
20000	`vegetables(tomato).`

の場合に food(X) を問い合わせすると、tomato,orange,apple の順で結果が出る。もし、orange が出力されたところで推論を止めれば、`food(X):-fruit(X)` と `fruit(orange)` のコストが減少し、次に food(X) を問い合わせしたときに、orange,apple,tomato の順で結果が出る。

3 MetaRoamer:転送先学習型メタ検索エンジン

図 1 に MetaRoamer の構成を示す。

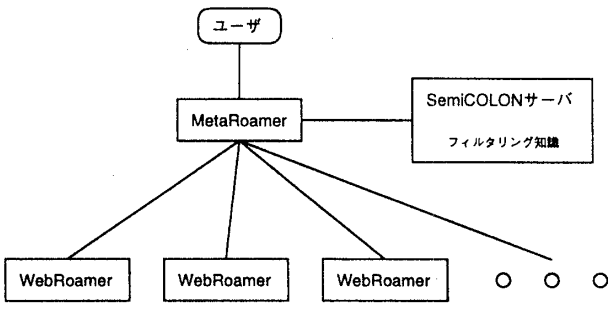

図 1: MetaRoamer の構成

3.1 WebRoamer

WebRoamer は MetaRoamer の転送先検索エンジンとして開発された小規模検索エンジンで、述語論理形式の検索式を解釈することができる。WebRoamer は手持ちデータの種類により 8 つ用意した。

3.2 SemiCOLON サーバ

SemiCOLON サーバは SemiCOLON の MetaRoamer 用インターフェースである。MetaRoamer から検索式を入力として受け取ると、その検索式に適した検

A Web Metasearch Engine Using Information Filtering Technique
Daishi Kato, Masayuki Numao
Department of Computer Science, Tokyo Institute of Technology

索エンジン名の順序を出力する。また、MetaRoamer からのユーザのクリック情報をもとにSemiCOLON のフィルタリング知識のコストを変化させる。

3.3 MetaRoamer

MetaRoamer はユーザとのインターフェースを担うメタ検索エンジンである。MetaRoamer の動作は、

i) ユーザから検索式を受けとる

ii) 検索式を各 WebRoamer に転送する

iii) SemiCOLON サーバに問い合わせ WebRoamer の順序を得る

iv) 各 WebRoamer から返ってきた検索結果を SemiCOLON サーバから得られた順序でユーザに提示する

である。

検索式に利用する述語は以下の8つである。

```
category(+CATEGORY)
type(+TYPE)
feature(+FEATURE)
word(+WORD,+COUNT,+MATCH_CASE)
two_words(+WORD1,+WORD2,+DISTANCE,+COUNT,+MATCH_CASE)
tag_count(+TAG,+COUNT)
word_inside_tag(+TAG,+WORD,+COUNT)
word_between_tag(+TAG,+WORD,+COUNT,+MATCH_CASE)
```

SemiCOLON サーバをこの検索式に対応させるため、フィルタリング知識として検索式を分解しそれぞれの文字列と検索エンジンを関連付ける述語を作成した。

述語論理形式の検索式をユーザに直接入力させるのは困難であるため、項目追加形式インターフェースを用意した。このインターフェースにより、ユーザはマウスのクリックとフィールドへの文字入力だけで検索式を作ることが可能である。図2に項目追加形式の例を示す。

4 実験

MetaRoamer を Web(http://numao-www.cs.titech.ac.jp/~daishi/metaroamer/) で公開し、およそ1ヶ月で 500 件のデータを収集した。収集した 500 件のデータのうち 50 件を試験データとし、残りの 450 件を訓練データとして、SemiCOLON の性能評価を行なった。図3は、訓練データを加えていくときの試験データの正答率 (被験者が示した通りの検索エンジンを初めに出力した割合) を表している。横軸は加えた訓練データの個数で縦軸は正答率である。

また、試験データの選択に依存しないようにするため、10-fold クロスバリデーションにより計算した結果、訓練データをすべて加えたときの正答率は平均59.8%であった。

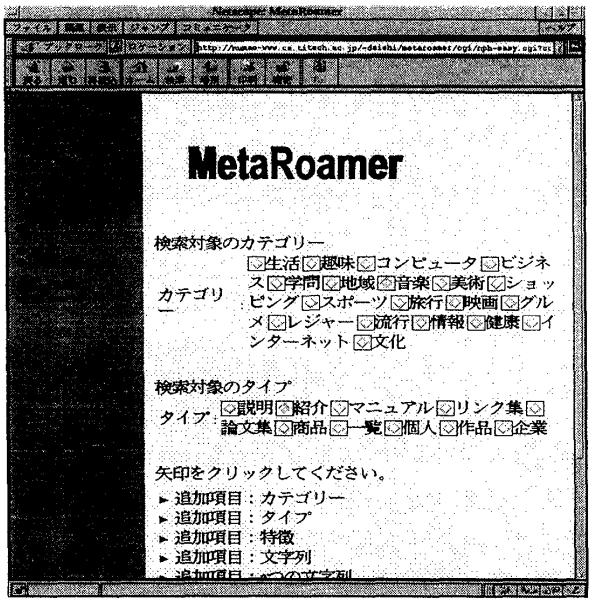

図 2: 項目追加形式の例

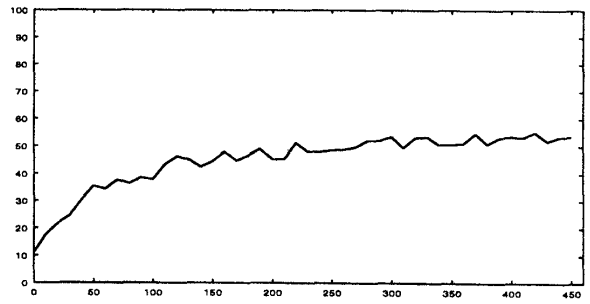

図 3: 実験結果

5 まとめ

実験結果から分かることは、8種類の検索エンジンから一つの検索エンジンを選ぶ確率 1/8 が SemiCOLON により 60%近くまで向上したということである。学習効果の意味では高い数値ではないが、インクリメンタルな学習が可能なことや不要なものを排除するフィルタリングの意味では評価できる数値である。

MetaRoamer は SavvySearch[1]のフィルタリング部と機能が非常に似ているが、SavvySearch は文字列と検索エンジンを一元に関連付けているのに対し、MetaRoamer は SemiCOLON のフィルタリング知識を変更するだけで複雑な検索式に対応でき、より高次なフィルタリングが可能であることが特徴である。

参考文献

[1] Howe, A. E. and Dreilinger, D.: SavvySearch A Metasearch Engine That Learns Which Search Engines to Query, *AI Magazine*, Vol. 18 (1997), 19–25.