D-002

スポンサードサーチにおける近似辞書検索アルゴリズム

Approximate Dictionary Lookup Algorithms for Efficient Sponsored Search

潘　献宇†　　福岡　正教†　　成　凱‡
Xianyu Pan　Masanori Fukuoka　Kai Cheng

1．はじめに

スポンサードサーチ(Sponsored Search)とは，Web 検索を行う際に，検索ボックスに入力されたキーワードと関連する広告をスポンサーリンク等の形で検索結果に表示させるオンライン広告の仕組みである。検索エンジン側においてスポンサードサーチを実現するために，まず，事前に登録された広告から辞書検索を行い，検索条件と一致する広告候補を検出する。そして，検索結果に一度表示される広告の数が限られているため，検出された候補結果に対し，更にほかの基準で厳選し最も適切な広告リンクを検索結果とともに表示させる。例えば，検出の広告候補から，オークションなどのメカニズムにより，実際に表示される広告を競争的に選出する方式が一般的である。

スポンサードサーチにおける辞書検索では，再見率より適合率（全検索結果に対しての検索要求を満たす検索結果の割合）を重視するため，精度の高い検索手法を用いることが一般的である，例えば，部分一致検索(broad match)，完全一致検索(exact match)，フレーズ一致検索(phrase match) [1]。一方，広告掲載依頼者にとって，適切なキーワードを登録することが困難である問題点が指摘されており，より柔軟なキーワード登録方法とそれに適した広告検索の仕組みが必要と思われる。

本論文では，我々はスポンサードサーチのために，キーワードを必須語，不要語，重み付き検索語等複数のカテゴリに指定できる新しいキーワード登録方式を提案し，それに適した近似的辞書検索手法を開発することを目的とする。これによってスポンサードサーチにおける辞書検索の精度を保ちながら，広告掲載依頼を改善する。

2．スポンサードサーチにおけるキーワード登録

2.1 キーワード登録と広告検索方式

広告主（広告掲載を依頼する側）は検索サービス運営会社に広告掲載を依頼する際に，掲載予定の広告に適した検索語をキーワードとして登録する必要がある。登録したキーワードで検索された際に，広告主の広告が検索結果画面に表示される。キーワードの登録は，どのような検索がされたときに広告が表示されるかを決定する[3]。広告主は商品やサービスに関連するキーワードがわかればそれを入力すればよいが，どんなキーワードを登録すればよいか分からない場合，広告テキスト入力時に入力した URL の Web ページを分析した結果から関連すると考えられるキーワードからを選択することができる。

登録したキーワードだけではなく，適合広告検索の方式によっても表示される広告が異なる。主な検索方式は以下のものがある。

部分一致検索：登録したキーワードと検索キーワードが完全に一致しなくても，他の単語が含まれる場合も広告が表示される。より広範囲のユーザーに広告を表示できる。

完全一致検索：登録したキーワードと検索キーワードが完全に一致する場合のみ広告が表示される。より正確にターゲットを絞ったユーザーに広告を表示できる。

フレーズ一致検索：登録したキーワードと検索キーワードの順序が同じであれば，他の単語が含まれる場合も広告が表示される。

カテゴリ	キーワード入力
必須キーワード	・ 国産牛肉 ・ 和牛
不要キーワード	・ 安い ・ 輸入
順位付きキーワード	1. 黒毛 2. 近江牛 3. 高級 4. 柔らかい

図 1　多様なキーワード登録

2.2 多様なキーワード登録方式

キーワード登録をより柔軟に行えるために，我々は，以下のような３種類キーワードが指定できるような方式を考案する。

（1）必須キーワード（Must Keywords）：ここに登録されたキーワードは，検索条件に最低一つ含まれる必要がある。いずれも含まない場合は，広告を表示させない。

（2）不要キーワード(Stop Keywords)：ここに登録されたキーワードは検索条件に含まれてはいけない。一つでも含まれた場合は，広告を表示させない。

（3）重み付きキーワード（Weighted Keywords）：このカテゴリのキーワードが検索条件に現れると，該当する重みを加算する。重みが大きいほど，適合するキーワード数が多いほど，広告を表示させる可能性が高い。ただ，キーワードごとに重みの具体値を付けることが難しいので，我々は，重みの具体値を指定せず，キーワードに優先順位をつけるだけで，アルゴリズムは自動的に重みに変換する方式を考える。

2.3 キーワード種別に応じる広告検索

上記のキーワード登録方式に適した広告検索を実現するために，図 2 に示すような多段階検索を用いる。登録されたキーワードは，必須キーワード，不要キーワード及び重み付きキーワードのカテゴリに応じて，それぞれ Must 辞書，Stop 辞書，重み付き辞書にまとめる。辞書エントリーは以下のよう形式をもつ。

<キーワード，関連広告リスト>

検索語と**キーワード**と照合して一致した場合，関連広告リストを調べることできるようになる。また，重み付きキーワードの場合は，関連広告は重みの降順で並べられてい

† 九州産業大学大学院　情報科学研究科
‡ 九州産業大学　情報科学部

るとする。

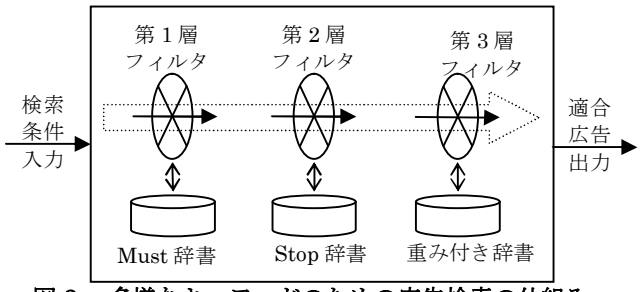

図 2　多様なキーワードのための広告検索の仕組み

検索語 W が入力されたとき，上記の枠組みを用いて辞書検索を行う。まず，第 1 層フィルタにおいて Must 辞書を検索し，適合する広告リストを Must(W)と表す。そして，第 2 層フィルタにおいて，Stop 辞書を検索し，適合する広告リストを Stop(W)とする。従って，次のフィルタにかけるべき候補広告は Must(W) – Stop(W) になる。次に，第 3 層フィルタに進み，Must(W) – Stop(W)にある候補広告の優先順位を決める。ここで重み付き辞書を用いて，近似辞書検索を行う。詳細について，後ほど詳しく説明する。

複数の検索語が与えられた場合は，上記と同じ手順でそれぞれ辞書検索を行う。得られた結果を統合し重みのもっとも大きなトップ K 個の広告を出力する。

3．近似辞書検索アルゴリズム

3.1 優先順位から重みへの変換

提案方式では，キーワードごとに重みを明示的に付けず，優先順位のみ指定すればよい。しかし，広告主の重み設定の負担を軽減できる一方，キーワードに対する各候補広告の重みの値がわからないため，候補広告表示の優先順位を決めることができない。

この問題を解決するため，ここで優先順位から重みの具体値を計算する方法を考える。まず，一つの候補広告に n 個のキーワードがあるとする。キーワードは重みの高い順で並んでいる。すると，i 番目のキーワードの重みは

$$w_i = \frac{n-i}{\sum_{i=0}^{n-1}(n-i)} \times 10 = \frac{20(n-i)}{n(n+1)}$$

このように得られた重みは[0, 10]の間にある。キーワード数 n によって優先順位が同位であっても重みに差がある。例えば，表 1 では第 1 位の重みは n=3, 9, 20 の時，それぞれ 5.0，2.0，1.0 となっている。つまり，キーワード数が少ないほど，同じ順位にあるキーワードの重みが高い。これは一見不公正に見えるが，実はそうでもない。理由は，登録キーワード数が多いほど，検索によってヒットされるチャンスが高いので，最終的に広告が表示される機会は平等といえる。

i	0	1	2	3	4	5	6	7	8
n=3	5.0	3.3	1.7						
n=9	2.0	1.8	1.6	1.3	1.1	0.9	0.7	0.4	0.2
n=20	1.0	0.9	0.9	0.8	0.8	0.7	0.7	0.6	0.6

表 1　優先順位から重みへ変換の例

3.2 近似辞書検索の実現

重み付き自称検索は以下のような不完全リストに対する Top-K 検索として扱うことができる。つまり，「与えられた m 個の順序付きリストから総合的重みが最も大きな K 個のデータを探し出す。」ここで，m 個の順序付きリストは，検索利用者が提出した検索条件に検索 Must 検索，Stop 検索を通して，重み付き検索で得られた m 個の結果（広告リスト）のことである。図 3 は m=3 の例を示している。複数のリストにおけるある広告の重みを統合するために，合計関数を用いる。例えば，w(a3)=2.00+5.00=7.00

順位	List 1		List 2		List 3	
	広告	重み	広告	重み	広告	重み
0	a_4	3.30	a_0	2.90	a_3	5.00
1	a_2	2.70	a_7	2.40	a_1	3.30
2	a_3	2.00	a_4	1.90	a_7	1.70
3	a_5	1.30	a_8	1.40		
4	a_6	0.70	a_2	1.00		
5			a_5	0.50		

図 3　不完全リストに対する Top-K 検索

Top-K 検索を効率的に行うために，FA，TA，BPA 等の有名なアルゴリズムが知られている[2]。しかし，これらのアルゴリズムはほとんど完全リストを対象として行われている。長さの異なる複数の不完全リストはうまく適用できるかは不明である。そこで我々は次のようなアルゴリズムを用いる。

1. m 個リストに対し先頭から同時調べていく。これまで触れたことのない新たなデータ d があるリストに現れた場合は，データ d の存在をほかのリストにも確認し，存在する場合は，そこの順位と重みを取得したうえ d の重みの合計を求める。
2. ここまで調べたデータのうち，総合重みの最も大きな k 個を集合 Y として保持する。さらにここで調べた順位とそこの重みの合計も記録しておく。
3. リスト i のこれまで調べたことのある順位が 1 位から連続している最大の順位をベスト順位 bp_i とし，そこのデータの重みを $w_i(bp_i)$ とする。さらに $\lambda = \sum_{i=1}^{m} w_i(bp_i)$。Y の要素のうち総合重みが λ より低いものがなければ，アルゴリズムが終了。そうでなければ，1 へ戻って繰り返す。

4．終わりに

本論文では我々はスポンサードサーチにおける新しいキーワード設定方法とそれに適した近似辞書検索の方式を提案した。優先順位から重みへ変換の方法や不完全リストに対する Top-K 検索についても検討を行った。紙面の制限で詳しい結果を省略している。

参考文献

[1]. A. C. König, K. Church, M. Markov. A Data Structure for Sponsored Search. In Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE 2009), pp.90-101, 2009

[2]. R. Akbarinia, E. Pacitti and P. Valduriez. Best Position Algorithms for Top-k Queries, In Proceedings of VLDB 2007, pp. 495-506, 2007

[3]. Google アドワーズご利用ガイド〈基礎編〉，Google 社，2009 年 7 月